MARRIAGE & FAMILY REVIEW
2024, VOL. 60, NO. 7, 395–433
https://doi.org/10.1080/01494929.2024.2392672

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Resolving the Contradictory Conclusions from Three Reviews of Controlled Longitudinal Studies of Physical Punishment: A Meta-Analysis

Robert E. Larzelere[a], Marjorie Lindner Gunnoe[b], Joshua Pritsker[c] and Christopher J. Ferguson[d]

[a]Oklahoma State University, Stillwater, Oklahoma; [b]Calvin University, Grand Rapids, Michigan; [c]Purdue University, West Lafayette, Indiana; [d]Stetson University, DeLand, Florida

ABSTRACT

Three literature reviews of controlled longitudinal studies of child outcomes of customary physical punishment have arrived at contradictory conclusions. We attempt to explain the contradiction using meta-analyses based on two types of change-scores and two sensitivity tests. We hypothesized that studies employing standard ANCOVA-type longitudinal analyses would suggest harmful-looking (but trivial) effects, and studies employing simple difference scores would suggest beneficial-looking (but trivial) effects of customary spanking on four child outcomes. We hypothesized that age (18 mos – 11 yrs) would moderate these associations. We hypothesized that spanking to enforce timeout in randomized studies of clinically defiant young children would predict large improvements in children's cooperation with timeout and parental commands. All hypotheses were supported. Regardless of statistical method, customary spanking explained less than 1% of the remaining variance in each child outcome (after controlling for baseline adjustment). The oft-reported harmful-looking outcomes of customary physical punishment in ANCOVA-type analyses are likely due to residual confounding. Various methodological problems and needed innovations in parental discipline research are discussed. Given the seeming near-zero effect of customary spanking, and the large beneficial-looking effects of spanking to enforce time-out in clinic-based intervention programs, blanket anti-spanking injunctions are discouraged.

KEYWORDS

causal validity; longitudinal; meta-analysis; spanking

Introduction

Physical punishment is clearly correlated with adverse child outcomes. Bivariate associations between physical punishment and a variety of negative outcomes have been demonstrated in two meta-analyses (MAs) by

CONTACT Robert E. Larzelere robert.larzelere@okstate.edu Department of Human Development & Family Science, 233 NRD Bldg., Oklahoma State University, Stillwater, OK 74078, USA.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/01494929.2024.2392672.

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

Gershoff (Gershoff, 2002; Gershoff & Grogan-Kaylor, 2016) and one by Paolucci and Violetti (2004). Gershoff's second MA was emphasized by the American Psychological Association (Gershoff et al., 2018) and the American Academy of Pediatrics (Sege et al., 2018) as part of the scientific basis for their opposition to all disciplinary spanking.

Whether physical punishment *causes* adverse child outcomes is less clear (Gershoff et al., 2018; Larzelere et al., 2019). To address this question, researchers must (at minimum) conduct longitudinal studies that control for children's preexisting differences at "baseline" (i.e., before or contemporaneous with the experience of physical punishment). The corpus of longitudinal studies that control for baseline outcome measurements has now been summarized in three published literature reviews (Ferguson, 2013; Heilmann et al., 2021; Larzelere, Gunnoe, et al., 2018), with the authors reaching discordant conclusions. The overarching goal of the present paper was to synthesize these three reviews, clarifying the causal relationship between spanking and multiple child outcomes and discussing implications for research, policy, and practice.

Three relevant reviews

Ferguson (2013)

The first review of statistically controlled longitudinal studies was a meta-analysis. Ferguson used partial *r*s to compute the average effects of physical punishment on three child outcomes (internalizing problems, externalizing problems, cognitive performance) beyond what was predicted by baseline scores on these outcomes. Like most meta-analyses of physical punishment since 2004, Ferguson specifically excluded abuse. Unlike earlier meta-analyses, Ferguson included only longitudinal studies and conducted separate analyses for spanking (specifically) and corporal punishment (more broadly defined). After controlling for initial differences in child adjustment, all of Ferguson's mean effect sizes were in a harmful-looking direction. However, the effect sizes were all *trivial* or small, ranging from partial *r*s of .07 to .10 for spanking and from .08 to .11 for corporal punishment (i.e., explaining only 0.5% to 1.2% of the remaining variance in subsequent child outcomes). Accordingly, Ferguson concluded that "the impact of spanking and [corporal punishment] on the negative outcomes evaluated here (externalizing, internalizing behavior and low cognitive performance) are minimal" (Ferguson, 2013, p. 196).

Larzelere, Gunnoe, et al., (2018)

The second review of statistically controlled longitudinal studies was a meta-analysis focusing exclusively on externalizing problems. As part of a special issue of *Child Development* on meta-analytic methods in

developmental science, Larzelere and colleagues introduced several techniques to improve the utility of meta-analyses of controlled longitudinal studies for distinguishing true causal effects from correlational associations and statistical artifacts. Of the three relevant reviews, this one employed the strictest inclusion criteria. Children needed to be at least 18 months and younger than 12 years. Physical discipline was limited to openhanded or customary spanking.

The primary focus of the Larzelere et al. meta-analysis (MA) was the comparison of two equally valid methods for analyzing change. The researchers called the two methods the "beta method" and the "slope method." Similar to Ferguson's partial *r* method, the *beta* method predicts outcomes at Time 2 (T2) controlling statistically for outcome scores at Time 1 (T1). The *slope* method predicts simple change scores (T2 – T1). (More on this in the Method section.) Effect sizes for both methods can be calculated from the correlations among three variables: T1 spanking and the child outcome scores at T1 and at T2.

The estimates obtained from these two change-score methods are often biased in opposite directions, when applied to longitudinal investigations of corrective actions. (By *corrective action*, we mean an action applied to correct a particular problem. For example, chemotherapy is an action applied to correct cancer; spanking is an action applied to correct misbehavior.)

The beta method is typically biased *against* corrective actions in longitudinal analyses (Berry & Willoughby, 2017; Hamaker et al., 2015; Larzelere, Lin, et al., 2018). Unless a researcher controls perfectly for all factors associated with the implementer's choice to use a corrective action (e.g., via randomization or a regression discontinuity design: Van Breukelen, 2013), a corrective action will typically be positively associated with subsequent symptoms related to the preexisting problem that prompted its use. (For example, those who have undergone chemotherapy will manifest more subsequent cancer than healthy adults who had no reason for chemotherapy. Likewise, children who elicited spanking will manifest higher rates of subsequent problem behavior than children who presented no reason for spanking in the first place.) This bias in beta-type analyses has been demonstrated against many corrective actions, including out-of-home placements (Berger et al., 2009); psychotherapy, *Ritalin*, other medications, spanking, and other parental disciplinary responses (Larzelere, Cox, & Smith, 2010; Larzelere, Ferrer, et al., 2010; Larzelere, Lin, et al., 2018).

In contrast, the slope method may be biased *in favor* of corrective actions. Because corrective actions are more likely to be applied when the rate or severity of a problem is high, natural regression to the mean can make the corrective action appear to be more helpful than it actually is (Larzelere, Lin, et al., 2018, Figure 3, p. 248).

Researchers can be more confident in the true direction of an "effect" that is robust across *both* methods (Angrist & Pischke, 2009; Duncan et al., 2014). If a researcher obtains a small effect in the *expected* direction (based on a technique's known bias), the effect may be simply a statistical artifact of the technique employed. However, if a researcher obtains a small effect in the *opposite* direction (i.e., the effect was large enough to overcome the technique's known statistical bias), the researcher can have more confidence that the effect is a true one.

Results from the MA by Larzelere et al. were *not* robust across the beta and slope methods but suggested "a small risk or a small benefit of spanking, *depending on the adjustment method*" (Larzelere, Gunnoe, et al., 2018, p. 2038). Put another way, both methods yielded a small significant finding in accordance with that method's known statistical bias that was *not* strong enough to overcome the opposite bias in the alternative method. This prompted Larzelere and colleagues to propose that there may be *no* consistent average causal effect of customary spanking on subsequent externalizing problems for children 18 months through age 11.

Heilmann et al. (2021)

The third review of statistically controlled longitudinal studies was a "narrative" or "box-score" review of the effects of physical discipline on nine child outcomes (Heilmann et al., 2021). In contrast to meta-analyses, which provide an average effect size, a narrative review simply categorizes studies as reporting harmful-looking, beneficial-looking, mixed, or non-significant outcomes – *without* quantifying the magnitude of those effects.

Other differences between Heilmann et al.'s narrative review and the two relevant MAs are that Heilmann et al. included a larger age range (birth to 18 years old), defined physical punishment more broadly (only excluding "severe assaults," p. 356), included some studies where the first measure of child adjustment occurred *after* the measure of physical punishment (i.e., was not truly a *baseline* measure), included multiple publications from the same data, and examined a wider range of outcomes. These more liberal inclusion criteria afforded Heilmann et al. many more studies to analyze ($n = 69$), compared to Ferguson's ($n = 24$) or Larzelere, Gunnoe, and Ferguson's ($n = 17$) eligible studies.

Most of Heilmann et al.'s main effects were harmful-looking (69% of studies). Only 1% had more beneficial-looking than harmful-looking main effects. Commenting on these findings, Heilmann and her colleagues (2021) concluded that their review "documented compelling evidence that physical punishment is harmful to children's development and wellbeing and … [that] all countries should … promote their wellbeing by prohibiting physical punishment in all forms and all settings" (p. 361).

Heilman's et al.'s demonstration of mostly harmful-looking effects is not surprising because most of the studies included in this narrative review employed a version of the beta method. Most beta-type coefficients reported in the two relevant meta-analyses (Ferguson, 2013; Larzelere, Gunnoe, et al., 2018) were also in a harmful-looking direction. Evidence that this pattern may be attributable to the aforementioned bias in the beta method is consistent with results from explicit attempts to *overcome* this bias. Such attempts have resulted in harmful-looking effects of spanking becoming non-significant (Berry & Willoughby, 2017, pp. 1198–1202; Pritsker, 2021, pp. 2598–2599: full sample) or beneficial-looking (Berry & Willoughby, 2017, online Supporting Information, pp. 4–6; Pritsker, 2021, pp. 2598–2599: limited sample).

Goals and hypotheses

As already stated, the *overarching* goal of this report was to reconcile the contradictory conclusions from these three most relevant reviews and discuss implications. Thus, we limited the present investigation to studies available at the time of the Heilmann et al. review. Although we considered conducting the most up-to-date MA possible, we wanted to ensure that readers did not attribute any differences in recommendations across publications to additional studies published after Heilmann et al. went to press.

We also had three *specific* goals. Specific goal #1 was the basis for the primary analyses. Inclusion of specific goals #2 and #3 was necessary to evaluate Heilmann et al. (2021) recommendation that all countries should prohibit physical punishment in all settings *and forms*.

Specific goal #1

The first specific goal was to estimate true average effect sizes for the causal impact of *customary* physical punishment on multiple child outcomes, by including analyses that are less biased than the beta method. To accomplish specific goal #1, we used the same basic protocol employed in the most recent MA (Larzelere, Gunnoe, et al., 2018). We used a similar age range (1.5 to 11.9 years) because this is the group for whom the efficacy of spanking is being debated. (We know of no academic researchers who "permit" physical discipline with children younger than 18 months or older than age 11.) We computed meta-analytic summary statistics using the beta and slope methods and subjected these analyses to additional sensitivity tests intended to distinguish true vs. artifactual effects.

We also expanded the 2018 MA in three ways. First, we considered for inclusion all studies identified in the three most relevant reviews of controlled longitudinal studies (Ferguson, 2013; Heilmann et al., 2021;

Larzelere, Gunnoe, et al., 2018). Second, we examined the four most commonly investigated child outcomes (not just externalizing problems). Third, we included *multiple* sequential intervals from datasets with more than one qualifying interval (The 2018 MA included only the *first* interval within the requisite age range.). Including multiple intervals was appropriate because effect sizes sometimes vary considerably across intervals (e.g., Coley et al., 2014), which may help explain some contradictory conclusions across the three reviews. Including all qualifying intervals seemed more sound, and it provided more data to test whether the effect sizes for customary spanking vary by the age at which children were spanked.

Specific goal #2

Our second specific goal was to identify situational (e.g., child age) and methodological (e.g., same-source bias) factors associated with better- or worse-looking outcomes than the average effects of customary physical punishment. To accomplish specific goal #2, we conducted moderator analyses to identity factors associated with the seeming effects of customary spanking.

Specific goal #3

Our third specific goal was to estimate unbiased average effect sizes for *back-up spanking* (i.e., two swats to a child's bottom only for refusing to cooperate with timeout) in randomized clinical trials. From the 1960s to the mid-1990s, most empirically-supported clinical treatments for disruptive behavior problems taught parents to administer two swats to the buttocks to enforce cooperation with timeout (Reitman & McMahon, 2013). The effectiveness of the "two-swat" spank that was taught in Forehand and McMahon (1981) training program was evaluated in a series of randomized studies conducted by Roberts and his colleagues (Bean & Roberts, 1981; Day & Roberts, 1983; Roberts, 1988; Roberts & Powers, 1990). Of all available studies of spanking, this set of studies best approximates the methods used to evaluate treatment efficacy in the medical field. Like clinical trials required for prescription medications, these four studies focused on participants with a precise presenting problem, randomly assigned participants to one of two or more treatment conditions, and prescribed a precise dosage/administration of treatment. To oppose *all* spanking "in all forms" requires evidence that the *most appropriate use* of spanking is ineffective or harmful.

To accomplish specific goal #3, we conducted a separate analysis of these randomized trials. According to the Cochrane and Campbell Collaborations, systematic reviews of non-randomized studies should always

provide separate effect sizes from eligible randomized studies that are relevant to the research question, because they provide the least biased causal evidence (Methods Group of the Campbell Collaboration, 2019; Reeves et al., 2019).

Hypotheses

Corresponding to these goals are the following hypotheses.

Hypothesis 1: Results from our multi-method meta-analyses of customary spanking will be consistent with all three literature reviews. Specifically, we expect harmful-looking trivial mean effect sizes associated with the beta method, which reverse to beneficial-looking trivial mean effect sizes when re-analyzed with the slope method.

Hypothesis 2: Spanking will be associated with smaller harmful-looking effects for younger children within our requisite age range. This hypothesis is based on both relevant meta-analyses (Ferguson, 2013; Larzelere, Gunnoe, et al., 2018). Tests of other moderators were exploratory.

Hypothesis 3: Back-up spanking will be more effective than the control condition in randomized interventions with clinic-referred 2- to 6-year-olds. Compared to the control condition, we expect back-up spanking to be associated with faster cooperation with timeout and greater compliance with parental commands.

Method

To the extent possible, we followed "state-of-the-art" guidelines for improving the validity of causal inferences in meta-analyses of non-randomized studies (Methods Group of the Campbell Collaboration, 2019; Reeves et al., 2019; Wells et al., 2013). For example, these guidelines specify that meta-analysts should provide precise statements of their research question, how their "intervention" is defined, what the intervention is being compared with (e.g., the control condition, sometimes called the "comparator"), and the minimally acceptable research design.

Meta-analytic specifics

The first research question was "*What are the effects of customary spanking on common child outcomes?*" The four child outcomes we examined were externalizing problems (broadly defined to include aggression, antisocial behavior, defiance, and total behavior problems); internalizing problems; cognitive achievement; and prosocial behavior/social competence. Each outcome was represented by at least four datasets in the three most relevant reviews.

The intervention was spanking. Ideally, we would have employed the precise definition of spanking provided by Friedman and Schonberg (1996)

– "administered with an opened hand to the extremities or buttocks" (p. 853) – but few studies have limited their measure to openhanded swats. Following other meta-analyses, we therefore included studies of "customary" spanking based on parent or child report of usage or frequency of undefined terms such as "spanking," "physical punishment," or "corporal punishment." We *excluded* studies that added other terms that made the operational definition too broad (e.g., push, shove, yell) or too severe (hit, use object). The only exception to this decision rule involved phrases like "spank or slap" with children under the age of four because slapping a young child's hand would fit the Friedman and Schonberg (1996) definition of an opened hand to the extremities.

Similarly, we had to rely on a less-than-ideal comparator: little or no spanking for a limited period of time (most commonly, one week). Obviously, the absence of spanking for one week cannot be assumed to indicate no spanking ever, but meta-analysts can only work with the studies available. They should, however, clarify the precise intervention tested and what it was compared with, to prevent generalizing their conclusions beyond the primary contrasts made in the included studies.

To be eligible, studies had to use research designs that approximated valid causal inferences at least as well as longitudinal designs that controlled for initial scores on the outcome viable or a reasonable proxy thereof. Consistent with contemporary guidelines (Reeves et al., 2019), we also included studies with stronger designs (randomized or quasi-experimental designs) if they met our other criteria. We required that children in the studies be at least 18 months and younger than 12 years when they were being spanked (i.e., at the start of a qualifying interval). We *excluded* samples dominated by children with developmental disabilities or reported to child protective services. At least two authors independently coded each eligible study; discrepancies were resolved by consensus.

Literature search, coding, and computing individual effect sizes

We considered all studies in the three prior reviews of controlled longitudinal studies of physical punishment (Ferguson, 2013; Heilmann et al., 2021; Larzelere, Gunnoe, et al., 2018). Of the 69 studies in Heilmann et al., 38 were eligible for inclusion and 32 were deemed ineligible for reasons shown in Table 1 (One study was eligible for one outcome but ineligible for a second, due to the lack of baseline data for the second outcome.). Twenty-eight of these 38 eligible studies had been included in one or both of the previous meta-analyses (Ferguson, 2013; Larzelere, Gunnoe, et al., 2018). We also identified nine additional eligible studies from the previous two meta-analyses, including four randomized trials. Thus, the total number of studies meeting our eligibility criteria was 47.

Table 1. Reasons why studies in Heilmann et al. (2021) failed to meet the inclusion criteria herein.

Study	Exclusion Criterion[a]	Details
Akcinar & Baydar, 2016	Too broad	Restraint, taking toy away, threats, spank/hit (observed)
Ansari & Gershoff, 2016	Concurrent	Year-2 spanking on Year-2 outcomes
Bakoula et al., 2009	Late control	Initial scores on outcomes occurred after spanking
Beauchaine et al., 2005	Too broad	Slap/spank/hit, restrain, lead/drag to corner, 'bad' timeout, push away, argue/fight
Bugental et al., 2003	Too broad	Spank/slap, push/shove, throw object at child, grab
Callender et al., 2012	Too broad	Spank, grab, shake
Cuartas et al., 2020	Too young	9 to 26 mos-old at Time 1 (M age = 17.8 mos)
Derella et al., 2020	Severe	"Spank or hit" 5- to 15-yr-olds
Ellison et al., 2011	Poor control	Unhappiness deemed an insufficient proxy for internalizing
Font & Cage, 2018	Concurrent	Predicted initial outcome score from initial spanking & outcome slope from spanking slope; 8 to 14 yrs at Time 1
Foshee et al., 2005	Severe	Hit or spanked (M age = 14.5 yrs)
Grogan-Kaylor, 2004	Concurrent	Past-week spankings with past-3-months antisocial, using deviations from each family's mean scores'
Grogan-Kaylor, 2005a	Concurrent	Past-week spankings with past-3-months antisocial, using deviations from each family's mean scores
Grogan-Kaylor, 2005b	Concurrent	Past-week spankings with past-3-months antisocial, using deviations from each family's mean scores
Keyser et al., 2017	Too young	0 to 12-mos-old at Time 1
Lahey et al., 2008	Too young	0 to 11-mos-old at Time 1
Lee et al., 2014	Too young	9- to 36-mos-old at Time 1 (M age = 15 mos)
Ma & Grogan-Kaylor, 2017	Concurrent	Associations of past-month spanking at age 3 or 5 with outcome at the same age for past 2 or 6 months
Ma et al,. 2018a	No control	No control for initial or prior scores on outcome.
Ma et al., 2018b	Concurrent	Past-year spanks with outcomes past 2 or 6 months, using deviations from each family's mean scores.
Ma et al., 2020	Concurrent	Past-year spankings with outcomes past 2 or 6 months, using deviations from each family's mean scores
Neaverson et al., 2020	Severe	Included slapping 11-yr-olds & pulling hair or ears
Okuzono et al., 2017	Poor control	Temperament: inadequate control for later behavior problems
Olson et al., 2011	Too broad	Included grab, shake, and other actions
Piché et al. 2017	Severe	If either parent "hit child because difficult" (1 of 2 items)
Scott et al., 2014	No control	Controlled for demographics, not initial scores on outcome
Slack et al., 2004	Outcome	Outcome (later allegation of neglect) is not included herein
Slade & Wissow, 2004	Too young	M age = 1.2 yrs-old at Time 1
Wang & Kenny, 2014	Too broad	Slap, push, grab, shove during past 6 mos
Xing et al., 2011	Too broad	Unspecified 3 items from Conflict Tactics Scale (Assumed too broad)
Yoo & Huang, 2013	No control	Controlled for inept parenting, domestic violence, & maternal mental health, but not initial outcome scores
Yu et al., 2018	Too broad	Grabs, slaps, & guides child by punishment more than by reasoning

[a]This lists the first applicable exclusion criterion in our pre-determined order. Many of the studies in the list were ineligible for *multiple* reasons (e.g., measure of discipline too broad and children too young/old).

All 43 eligible longitudinal studies included maternal reports of physical discipline, and 5 included separate paternal reports. To maximize the homogeneity and sample sizes of each study's contribution, we based the focal analyses on maternal reports, except for one study that measured spanking compositely, averaging across parents (Baumrind et al., 2010). To minimize single-source bias (i.e., association attributable to the use of the same reporter for both spanking and the child outcome), we preferred child outcome data from a non-maternal source when available (e.g., we

used mothers' report of spanking to predict child or teacher report of child externalizing or standardized tests for cognitive outcomes).

Following Lipsey and Wilson (2001), we combined information from the same longitudinal cohort at the same ages when summarized in multiple publications. For example, 17 of the 47 eligible studies analyzed data from the Fragile Families longitudinal dataset. We permitted publications from the same longitudinal dataset to make separate contributions to the focal MAs only if they focused on different birth cohorts or different age groups.

The steps for obtaining or deriving parameter estimates for each contribution to our four focal meta-analyses (i.e., each *line* in Table 2, subsequently referred to as "intervals") were as follows:

1. If authors provided a correlation matrix, either within a publication or in response to an emailed request, we utilized the correlations provided by the authors. We requested correlation matrices from several authors and are grateful to those who supplied them.
2. If multiple publications provided a correlation matrix for the same dataset/same interval/same outcome variable, we *averaged* the author-provided correlations, unless there was a clear reason to prefer one matrix to another (e.g., a more representative and/or larger sample).
3. To increase the number of intervals for less frequently investigated outcomes (i.e., outcomes other than externalizing), we calculated the three relevant correlations for less common outcomes from publicly available datasets that had been used to investigate externalizing behavior problems in any qualifying study. We were able to do this for two public datasets used in qualifying studies (Gershoff et al., 2012; MacKenzie et al., 2013). Parameters in Table 2 representing our *own* analyses of an eligible dataset are indicated with a footnote.

These three steps yielded 25 distinct qualifying intervals from 16 datasets (7 datasets had two or three qualifying intervals) with the requisite correlation matrices for our focal meta-analysis of externalizing problems. For internalizing problems, we identified 7 intervals from 5 datasets. For cognitive competence, we included 9 intervals from 7 datasets, and for social competence, 6 intervals from 4 datasets.

Studies that were eligible for the focal analyses, but *not* included in them, are listed in Table A1 in the Appendix. These studies represent overlapping samples/intervals and studies for which we could not obtain the three correlations necessary to compute the parameters needed for the focal analyses. We have also indicated in Table A1 studies with some

Table 2. Cross-sectional, longitudinal, and stability correlations among spanking (x_1) and child outcome scores (y_1 & y_2) used to predict ANCOVA-type residualized gain scores (β – Method) and simple gain scores in the outcome (slope method: $r_{x_1(y_2-y_1)}$) from spanking.

Source	Age[a]	*N*	N_{part}[b]	$r_{x_1y_1}$	$r_{x_1y_2}$	$r_{y_1y_2}$	*β*	$r_{x_1(y_2-y_1)}$
Externalizing behavior problems								
Barnes et al. (2013)	4.1	495	495	.07	.19***	.48***	**.16*** **	**.12*** **
Baumrind et al. (2010)	4.5	87	87	.35***	.01	.29**	**−.11**	**−.29****
Berlin et al. (2009)	2.0	1765	1765	.15***	.12***	.52***	**.04***	**−.03**
Coley et al. (2014, Int1)	3.4	501	270	.19***	−.05	.33***	**−.12***	**−.21*** **
Coley et al. (2014, Int2)	4.8	427	231	.12*	.28	.37	**.24*** **	**.14***
Ellison et al. (2011)	3.0	433	433	.22***	.14**	.25***	**.09^**	**−.06**
Fragile Families (FF Int1)[c]	3.0	2475	1471	.20***	.18***	.55***	**.08*** **	**−.02**
MacKenzie et al. (2013) FF Int2[d]	5.3	1688	1004	.21***	.15***	.47***	**.05***	**−.06***
Gershoff et al. (2012)	6.2	10492	10492	.15***	.15***	.51***	**.08*** **	**.00**
Gershoff et al. (2016, Int1)	3.6	2042	1031	.25***	.17***	.50***	**.05^**	**−.08****
Gershoff et al. (2016, Int2)	4.1	2001	1011	.28***	.22***	.47***	**.09*** **	**−.06^**
Gunnoe and Mariner (1997)	7.8	1027	1027	.15***	.15***	.12***	**.13*** **	**.00**
Lansford et al. (2011, Pitt Int1)	10.0	216	109	.29***	.29***	.61***	**.12***	**−.00**
Lansford et al. (2011, Pitt Int2)	11.0	214	107	.23***	.23***	.65***	**.08**	**.00**
Lansford et al. (2011 CDP Int1)	6.0	465	161	.15**	.15**	.57***	**.07^**	**.00**
Lansford et al. (2011 CDP Int2)	7.0	442	154	.08^	.10*	.61***	**.05**	**.02**
Lansford et al. (2011 CDP Int3)	8.0	433	150	.16***	.19***	.63***	**.09***	**.04**
Larzelere et al. (2010 Int1)	4.9	1464	488	.28***	.20***	.51***	**.07**	**−.07^**
Larzelere et al. (2010 Int2)	6.8	1464	488	.26***	.21***	.56***	**.07^**	**−.05**
Larzelere et al. (2010 Int3)	8.8	1464	488	.32***	.25***	.57***	**.08^**	**−.07^**
Mendez et al. (2016)	2.5	186	186	.30***	.21**	.60***	**.03**	**−.10**
Mulvaney and Mebert (2007)	3.8	979	979	.23***	.22***	.57***	**.09*** **	**−.01**
O'Gara et al. (2020, Int1)	4.4	248	138	.26***	.26***	.73***	**.08**	**.00**
O'Gara et al. (2020, Int2)	5.4	197	110	.37***	.36***	.74***	**.10**	**−.01**
Straus et al. (1997)[e]	7.5	785	785	.27***	.27***	.49***	**.15*** **	**.00**
Total N, meta-analytic *r*'s & β[f]			23660	.21***	.18***	.51***	**.08*** **	**−.03***
I^2(95% prediction intervals in footnote[g])				71.1%	57.7%	92.8%	31.1%	52.7%
Internalizing behavior problems								
Baumrind et al. (2010)	4.5	87	87	.11	.23*	.40***	**.19***	**.11**
Coley et al. (2014, Int1)	3.4	501	270	.06	−.13**	.41***	**−.16****	**−.17****
Coley et al. (2014, Int2)	4.8	427	231	.00	.22***	.23***	**.22*** **	**.18****
Fragile Families (FF Int1)[c]	3.0	2502	1510	.03	.07***	.43***	**.05****	**.03**
MacKenzie et al. (2013) FF Int2[h]	5.0	1644	992	.08**	.05*	.41***	**.02**	**−.03**
Mulvaney and Mebert (2007)	3.0	979	979	.18***	.09**	.44***	**.01**	**−.09**
Gershoff et al. (2012)[h]	6.2	7568	7568	.04***	.05***	.26***	**.04*** **	**.01**
Total N, meta-analytic *r*'s & β[f]			11637	.07**	.07**	.37***	**.04**	**−.01**
I^2(95% prediction intervals in footnote[g])				69.8%	70.0%	93.4%	72.0%	77.5%
Cognitive achievement								
Baumrind et al. (2010)	4.5	87	87	−.38***	−.19^	.39***	**−.05**	**.17**
Berlin et al. (2009)	2.1	1387	1387	−.08**	−.07**	.58***	**−.02**	**.01**
Gershoff et al. (2016, Int1)	3.6	2042	1031	−.03	−.03	.51***	**−.02**	**−.01**
Gershoff et al. (2016, Int2)	4.1	2001	1011	−.04^	−.04^	.60***	**−.02**	**.00**
Maguire-Jack et al. (2012, FF Int1)	3.0	1737	1000	−.03	.04^	.44***	**.06^**	**.07***
MacKenzie et al. (2013, FF Int2[d])	5.0	1280	737	−.04	−.05^	.65***	**−.02**	**−.01**
Straus and Paschall (2009, Cohort 1)	3.7	806	806	−.10**	−.12***	.35***	**−.09^**	**−.02**
Straus and Paschall (2009, Cohort 2)	7.7	704	704	−.16***	−.21***	.68***	**−.10****	**−.06**
Gershoff et al. (2012)[h]	6.2	8917	8917	−.12***	−.15***	.72***	**−.06*** **	**−.04*** **
Total N, meta-analytic *r*'s & β[f]			15680	−.08***	−.08**	.56***	**−.04***	**−.01**
I^2(95% prediction intervals in footnote[g])				74.0%	87.1%	98.3%	60.2%	56.3%
Social competence								
Altschul et al. (2016) FF Int1	3.0	3049	1980	.02	.02	.35***	**.02**	**.00**
MacKenzie et al. (2013) FF Int2[h]	5.0	1645	1069	.02	.00	.29***	**−.01**	**−.02**
Baumrind et al. (2010)	4.5	87	87	−.17	−.23*	.15	**−.21***	**−.05**
Gershoff et al. (2016, Int1)	3.6	2042	1031	−.08***	−.04^	.35***	**−.01**	**.04**

(*Continued*)

Table 2. Continued.

Source	Age[a]	*N*	N_{part}[b]	$r_{x_1y_1}$	$r_{x_1y_2}$	$r_{y_1y_2}$	β	$r_{x_1(y_2-y_1)}$
Gershoff et al. (2016, Int2)	4.1	2001	1011	−.03	−.08***	.33***	**−.07***	**−.04**
Gershoff et al. (2012)[h]	6.2	7549	7549	−.12***	−.14***	.37***	**−.10*****	**−.02**
Total N, meta-analytic *r*'s & β[f]			12727	−0.05	−.06	.34***	**−.04^**	**−.01**
I^2(95% prediction intervals in footnote[g])				89.2%	91.4%	62.1%	82.8%	0.00%

Note. Studies with two or three relevant intervals of the same children have multiple rows labeled Int1, Int2, & Int3 for Intervals 1, 2, and 3 from the youngest qualifying interval to the oldest qualifying interval. The three columns of correlations (near the middle) give the cross-sectional ($r_{x_1y_1}$) and longitudinal ($r_{x_1y_2}$) correlations between T1 spanking and the T1 and T2 outcome scores, respectively, followed by the stability correlations ($r_{y_1y_2}$) between the T1 and T2 outcome scores for each interval. The two right-hand columns give the standardized regression coefficients (β) predicting the T2 outcome from T1 spanking controlling for the T1 outcome score and the correlations ($r_{x_1(y_2-y_1)}$) between T1 spanking and the slope of the outcome score from T1 to T2 (i.e., T2 externalizing minus T1 externalizing).

[a]Mean child age at the first of two occasions for each row of data.

[b]Portion of the Total N for a particular study assigned to each interval (to avoid double counting children in meta-analytic calculations across intervals). The apportioned Ns for any given study sum to the actual N for the first interval.

[c]Mean of correlations from Altschul et al. (2016), Gromoske & Maguire-Jack (2012), Lee et al. (2013), and Maguire-Jack et al. (2012) for externalizing problems. Mean correlations from the two articles by Maguire-Jack and colleagues for internalizing problems.

[d]Based on attempted duplication of published results from Fragile Families data.

[e]Using correlations from attempted duplication by Larzelere et al. (2010).

[f]The summation rows present the Total N and the random-effects mean *r*'s and mean β's (Borenstein et al., 2009). The total *N* counted each dataset only once (i.e., the *N* from T1 to T2, not from T2 to T3 from datasets with two or three intervals represented by two or three rows).

[g]95% prediction intervals (Borenstein et al., 2009) for estimated range of "true" effects in last five columns: Externalizing problems: (.10, .33), (.09, .27), (.28, .69), (.03, .13), (−.11, .05); Internalizing problems: (-.06, .20), (−.07, .20), (.07, .61), (−.10, .18), (−.17, .15); Cognitive achievement: (−.20, .04), (−.26, .10), (.11, .82), (−.13, .05), (−.09, .08). Social competence: (−.27, .17), (−.31, .19), (.24, .43), (−.21, .13), (−.04, .01).

[h]Based on our analyses of the same dataset for an outcome not reported in the qualifying study.

^$p < .10$.

*$p < .05$.

**$p < .01$.

***$p < .001$.

unique features of potential interest – paternal spanking, less-frequent spanking (< once/week) and openhanded spanking – along with their most causally relevant effect sizes available (author-published or computed by our team when correlations were provided or estimated).

Multiple intervals from the same dataset

The inclusion of multiple intervals from the same dataset created two challenges for the estimation of *cross*-interval (but not *within*-interval) meta-analytic summary statistics. First, multiple intervals from the same children "double-counts" those children, inflating the sample size of the studies with multiple intervals. Second, effect sizes based on intervals from the same children are likely to be related to each other, whereas meta-analytic statistics assume independent samples.

To avoid double-counting children, we utilized attrition information to apportion each dataset's total sample size across the qualifying intervals. For example, if the correlation of T1 spanking to T2 externalizing was

based on $N = 100$, and the correlation of T2 spanking to T3 externalizing was based on only 75% of the 100 (due to attrition), we apportioned the original 100 participants across the two intervals using an algebraic equation (e.g., $1.0x + .75x = 1.75x = 100$; $x = 57$). We based meta-analytic summary statistics *across* intervals on the apportioned *n*s (57 and 43 for this hypothetical case). This weighting method reduces the influence of intervals with more attrition. To compute the *within*-interval meta-analytic summary statistics, we used the sample size of valid scores (e.g., 100 or 75 in the hypothetical case). Both the actual *N* for the interval and the apportioned N_{part} are provided in Table 2.

To address the issue of non-independent effect sizes, we utilized robust variance estimation (RVE: Hedges et al., 2010). RVE adjusts variance estimates to account for the fact that adjacent intervals in the same dataset are not independent of each other. Moeyaert et al. (2017) found that RVE performs better than other methods of controlling Type I error rates when the number of studies analyzed is not large. Thus, in line with expert recommendations, a small sample correction factor was utilized in all multi-interval analyses (Pustejovsky & Tipton, 2017; Tipton, 2015). The RVE method was implemented using the *metaphor* and *clubSandwich* R packages.

Smallest effect size of interest (SESOI)

As discussed in the Introduction, very small effect sizes should limit confidence in causal conclusions. Small effect sizes are more likely than larger ones to be due to residual confounding. Even *if* causally valid, small effect sizes represent a small difference across groups (e.g., explaining less than 1% of the variance in the outcome). Moreover, variations in the effect of any particular spanking are about as likely to be beneficial as detrimental, around a near-zero average effect. In large/high-powered studies – including meta-analyses – even tiny effects may become "statistically significant" and be mistakenly reported as unambiguously supporting a study's hypotheses, despite representing a trivial effect at best (often easily explained by systematic biases that remain). Although there is no standard cutoff for such false positive effects, Ferguson and Heene (2021) found that effect sizes below $r = .10$ (equivalent to Cohen's $d = .20$) had a very high false positivity rate despite statistical significance. They recommended *against* using effect sizes in this range as evidentiary support for a hypothesis. As such, we adopted $r = .10$ (ca. $\beta = .10$) as our minimum SESOI. Similar standards were used for "practical significance" in a recent study of predictors of child outcomes in 51 low- and middle-income countries ($r = .10$ or 2% of variance accounted for: Bornstein et al., 2023).

Analyses

Specific goal #1: Estimate unbiased effects of customary spanking

We computed two effect sizes (one for each method) for each qualifying interval listed in Table 2. We did this using *three central correlations*: T1 spanking with T1 outcome, T1 spanking with T2 outcome, and T1 outcome with T2 outcome. (Note: By "T1" and "T2" we mean the start and end of *each* interval listed in Table 2. For studies with multiple intervals, T2 for a prior interval is T1 for the next interval of the same sample of children.)

For the *beta method*, we used the three correlations to calculate each interval's standardized regression coefficient (β) for predicting a T2 child outcome from T1 spanking, controlling for initial scores on the outcome at T1. We then computed the overall weighted mean β for each of the four common child outcomes.

For the *slope method*, we used the same three central correlations to calculate the correlation of T1 spanking with the "slope" or change in child outcome (i.e., T2 outcome – T1 outcome) in a linear growth model for each interval. We then computed the overall weighted mean correlation of T1 spanking with the slope across intervals for each of the focal outcomes. Equations and sample syntax in SPSS and Mplus for calculating both the standardized β coefficient and the slope coefficient from these three correlations are given in the online Supporting Information for the meta-analysis by Larzelere, Gunnoe, et al. (2018).

Sensitivity tests. We also conducted two sensitivity tests related to specific goal #1. First, we examined whether controlling for *two* earlier externalizing scores (rather than just one) impacted the externalizing effect size for an interval. (Externalizing was the only outcome for which there was more than two studies with the necessary two intervals.)

Second, we subjected all significant mean effect sizes to the "backwards test" (Galton, 1886; a type of "falsification test": Pizer, 2016). This technique tests whether results replicate after reversing T1 and T2 for each interval (e.g., predicting T1 outcomes from T2 spanking controlling for T2 outcome scores). Residual confounding and statistical artifacts can often "work" backwards as well as forward in time, but actual causal effects only operate forward in time. If an effect produced in forward analyses is *not* replicated in reverse, researchers can be more confident that they have obtained a true causal effect than if the effect is replicated backwards in time (Larzelere, Gunnoe, et al., 2018).

Specific goal #2: Test moderators

We tested 13 potential moderators of the effects of customary spanking on externalizing problems using the CMA meta-analytic software (Borenstein et al., 2011). (Externalizing was the only outcome variable examined in

enough studies to permit moderator testing.) We tested 5 *continuous* moderators: child age, birth year, length of interval (from T1 to T2), percentage girls, and percentage ethnic minorities. We tested 8 categorical moderators. Following the advice of Fu et al. (2011), *categorical* moderators were tested only when we could delineate conceptually meaningful categories containing at least four qualifying intervals from Table 2 for each moderator category (e.g. for maternal education, we needed at least four studies/intervals with a "high school or less" category and four studies/intervals with a "post-secondary" category). This was possible for 8 categorical moderators for the externalizing outcome: sample type (nationally representative vs. low-SES national vs. local/regional), type of spanking measure (frequency/use vs. endorsement), independence of data sources (independent vs. partly or entirely from the same source), threshold for highest spanking score (3+ times per week vs. less frequent vs. other), specificity of time period for the spanking measure (past week/month vs. unspecified time), socio-economic status (low vs. average/above average), median maternal education (high school/less vs. post-secondary education), and percent of two-parent families (less than 68% vs. 68%-80% vs. 81% or more).

Specific goal #3: Compare back-up spanking with control condition in randomized trials

Using outcome data from a behavioral parent training program, we compared children who were prescribed *spanking to back-up time-out* with those in the *control-group* who were permitted to leave time-out at will (i.e., the "child-release" condition). Unfortunately, only two of the four randomized studies of back-up spanking included the child-release condition (Bean & Roberts, 1981; Roberts & Powers, 1990). To evaluate the effects of spanking in the other two studies (which compared back-up spanking with a brief room isolation), we needed an effect size for a control condition for comparison. We derived such an effect size by averaging the control group effect sizes from the two studies that *did* include the child-release control group and used their weighted average as the comparison for the back-up spanking condition for the two studies that *did not* include the child-release control group.

Concerning child outcomes, three of these four randomized studies of back-up spanking assessed *two* related measures of spanking effectiveness. The primary outcome was *improved cooperation with timeout*; the secondary outcome was *improved compliance to parental commands* from pretest to posttest. We calculated effect sizes for both outcomes and reported pretest and posttest scores on the secondary outcome to address previous critiques of these studies (more on this in the Discussion). The fourth randomized study (Day & Roberts, 1983) assessed *only* one measure of effectiveness (improved cooperation with parental commands).

Causal certainty of the evidence

To evaluate the causal certainty of the overall evidence, we employed the *Grades of Recommendation, Assessment, Development and Evaluation* (GRADE) method (Schunemann et al., 2019), using its four-point Certainty scale. GRADE users evaluate a set of studies by assigning an initial rating based on the strength of the research designs (e.g., randomized studies begin with a higher initial rating than non-randomized studies) and then adjusting up or down based on eight factors such as effect size magnitude and inconsistent results across the studies. GRADE is used by over 100 organizations, including the Cochrane Collaboration and the UK's National Institute for Health and Care Excellence (NICE), to assess the causal certainty of meta-analytic support for corrective actions by physicians. It is also applicable for corrective actions by parents. (See the online Supplementary Materials for more information about GRADE.)

Using GRADE, we estimated the certainty of the evidence for the average causal effect of *customary* spanking on *externalizing* problems (from Table 2), the most common child outcome in studies of spanking. However, customary spanking does not represent "physical punishment in all forms." Therefore, we also used GRADE to evaluate the causal certainty associated with the effectiveness of back-up spanking in clinical trials.

Results

The results varied by:

1. *Type of analysis:* We obtained different patterns of effects for beta vs. slope analyses.
2. *Child outcome variable:* The mean effect sizes for externalizing problems were slightly larger than the effect sizes for internalizing, cognitive achievement, and prosocial behavior/social competence.
3. *Type of spanking:* The effect sizes for back-up spanking were more beneficial-looking and much larger than effect sizes for customary spanking.

Effects of customary spanking

Specific goal #1 was to estimate true average effect sizes for the causal impact of customary spanking on multiple child outcomes, by contrasting effect sizes based on the beta method with effect sizes from analyses that are less biased than the beta method.

Beta method

The results obtained from the standard beta method (ANCOVA controlling for T1 outcome scores) replicated results from all three relevant literature

reviews. (See the next-to-last column of Table 2). Consistent with Heilmann et al. (2021) narrative review, 21 of the 47 *individual interval* betas (β's) were significant in a harmful-looking direction (i.e., more externalizing and internalizing problems than otherwise expected or less cognitive achievement and social competence than expected). Only 2 of the individual interval betas were significant in a beneficial-looking direction. Consistent with the two prior meta-analyses (Ferguson, 2013; Larzelere, Gunnoe, et al., 2018), the *mean meta-analytic* betas were significant for two of the four outcomes (externalizing problems: β = .08, $p < .001$; internalizing problems: β = .04, n.s.; cognitive achievement: β = −.04, $p < .05$; social competence: β = −.04, $p < .10$).

However, none of these effects met or exceeded our pre-specified Smallest Effect Size of Interest (SESOI) of $r = β = .10$ (Ferguson & Heene, 2021). Rather, they met Ferguson's (2013) definition of a *trivial* effect size, explaining less than 1% of the variance in each outcome after controlling for initial scores on that outcome. Specifically, spanking explained 0.64% of the remaining variance in externalizing problems, and 0.16% of the variance in each of the other three outcomes.

Results presented in Table 2 were generally replicated in the studies listed in Table A1. (Recall that these studies met our eligibility criteria but relied on the same dataset as studies in Table 2 or lacked the correlation matrices necessary for the main meta-analyses.) See Table A1 for the 26 additional studies on externalizing (only 4 addressed our other outcomes), the replicated results (e.g., β = .06, $p < .001$, for externalizing), and additional externalizing results from the few studies with unique features such as father data (β = .06, $p < .01$), openhanded spanking (β = .06, $p < .01$), or spanking less than once a week (β = .03, n.s.). Most of these overlapping results came from one dataset (Fragile Families).

Slope method

Use of the equally valid slope method produced mostly *nonsignificant* effects. (See the last column of Table 2.) The *significant individual* effects indicated that spanking was associated with improved outcome scores (six effects) more often than worsening outcomes (four effects). Concerning the *mean meta-analytic* effect sizes, only one was significant, and it was in the beneficial-looking direction: Spanking predicted significantly greater decreases in externalizing problems from T1 to T2, $r_{x1(y2\text{-}y1)} = -.03$, $p < .05$. As with the two significant meta-analytic betas, this slope effect size was also below our SESOI, explaining only 0.09% of the variance. The meta-analytic slope effect sizes for the other three outcomes were very near zero ($r_{x1(y2\text{-}y1)} = -.01$).

Again, results from the eight studies in Table A1 that provided the statistics necessary to calculate the slope statistic were consistent with the focal results presented in Table 2. Similar to the main analyses, the direction and magnitude for individual intervals and the meta-analytic effect size (mean $r_{x1(y2\text{-}y1)}$= −.04, p < .001) indicated that spanking predicted significant subsequent reductions in externalizing problems, albeit to a trivial degree. Studies with unique features (e.g., father data, openhanded, less frequent) that provided the correlation coefficients necessary to calculate a slope statistic were each limited to one dataset, resulting in non-significant slope correlations.

Sensitivity tests

Results of the autoregressive lag-2 (AR-2) analysis, controlling for externalizing scores at two prior time points (β = .07, p < .001; Supplementary Table S-1 online) were nearly identical to the ANCOVA (AR-1) results in Table 2 (β = .08, p < .001). For externalizing problems, the backwards test replicated the contradictory significant results in the main analyses (β = .05, p < .01; slope $r_{x1(y2\text{-}y1)}$= −.05, p < .05). The results for cognitive achievement were not significant (β = −.01, slope $r_{x1(y2\text{-}y1)}$= .01). There were insufficient studies to conduct backwards tests for internalizing and social competence.

Moderators

Specific goal #2 was to identify moderators of the average effect of customary physical punishment. The only significant moderator of *externalizing* (the only outcome with a sufficient number of eligible studies to test moderation) was child age (b = .008, ps < .05, see Figure 1). Consistent with our hypothesis, the effect of spanking on subsequent externalizing problems became more harmful-looking when using the beta method as children's age increased from 18 months (β = .05) to 11 years (β = .12). When using the slope method, the direction of the association reversed for younger children. Spanking was associated with slight reductions in externalizing behaviors at 18 months ($r_{x1(y2\text{-}y1)}$ = −.05) but slight increases at 11 years ($r_{x1(y2\text{-}y1)}$ = .03).

Effects of back-up spanking

Specific goal #3 was to estimate unbiased average effect sizes for the back-up spanking used in clinical trials, compared to the "child-release" control condition. Results from the four randomized studies of back-up spanking are shown in Table 3. Effect sizes for changes in *compliance to parental commands* from pretest to posttest (relative to letting children decide when

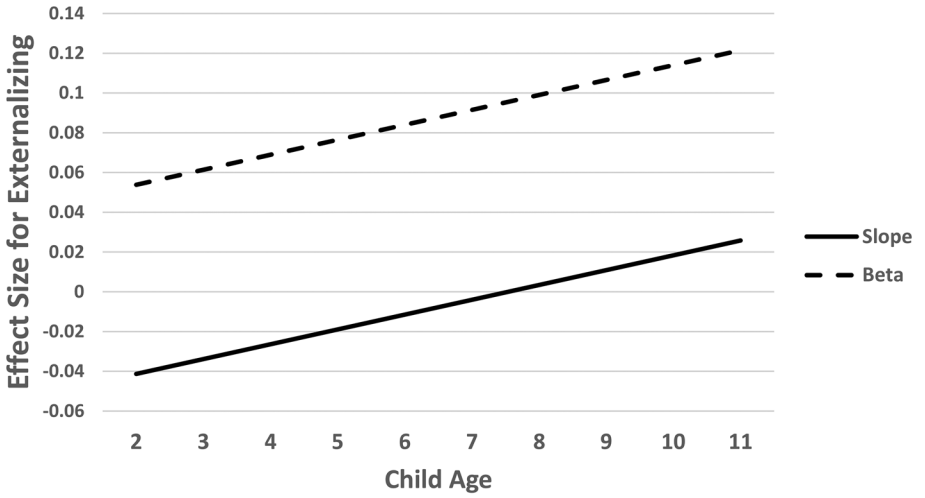

Figure 1. Effect size of the association of spanking with subsequent externalizing by child age and type of change-score analysis.

to end timeout) are provided in the third numerical column of the table ("Compliance *g*"). Effects sizes for how quickly children *cooperated with timeout* are presented in the fifth numerical column ("Timeout *g*"). The average of the two effect sizes is listed in the right-hand column ("Overall *g*").

In the two studies that directly compared back-up spanking with a child-release condition (top half of Table 3), spanking was more effective than the child-release control (Hedges' $g = 1.06$, $p < .01$). This meta-analytic effect size is larger than Cohen's (1988) guideline for a large effect in psychology ($g = .80$). Children in these two studies cooperated with timeout significantly more quickly when it was enforced with spanking than in the control condition in both studies ($g = 1.55$, $p < .001$). Spanking also resulted in significantly greater improvement in compliance to parental commands from pretest to posttest than the child-release control condition in one of the two studies ($g = 1.41$, $p < .01$, and $g = -.19$, n.s.). This pattern of the effectiveness of the spank back-up was replicated in two other randomized studies (bottom half of Table 3) that compared back-up spanking with an alternative back-up procedure (a brief room isolation, which is outside the scope of this meta-analysis).

Estimated causal certainty

As stated in the Method section, we used the GRADE method to evaluate the certainty of the causal evidence for the effect of (a) customary spanking on externalizing problems and (b) back-up spanking on compliance in behavioral training programs. Application of GRADE to *customary spanking* resulted in a rating of "Very Low" (the lowest rating on GRADE: Supplementary Table S-4).

Table 3. Effect sizes for back-up spanking on compliance with maternal commands and with timeout in randomized studies.

	Compliance with Maternal Commands			Compliance with Timeout		
Source	Pretest	Posttest	Compliance *g*	TO Criteria	Timeout *g*	Overall *g*
Studies with child release condition						
Bean and Roberts (1981, cell *n* = 8)						
Spank back-up for timeout	23.4%	77.9%	2.29**	6.5[a]		
Child release from timeout	23.3%	44.1%	.88*	16.6[a]		
Spank vs. child release (within study)			1.41**		1.61**	1.51**
Roberts and Powers (1990, cell *n* = 9)						
Spank back-up	18.0%	56.9%	1.25*	0%[b]		
Child release	23.9%	67.9%	1.44**	56%[b]		
Spank vs. child release (within study)			–.19		1.51*	.66
Meta-analytic weighted mean *g* (for above 2 studies)			**.56**		**1.55****	**1.06****
Studies with alternative comparison condition						
Day and Roberts (1983, cell *n* = 8)						
Spank back-up	21.3%	74.2%	3.31***			
Spank vs. child release (weighted *M*)[c]			2.13**			2.13**
Roberts (1988, cell *n* = 9)						
Spank back-up	30.7%	71.9%	1.52**	1.5[a] 11%[b]	3.47*** 1.02	
Spank vs. child release (weighted *M*)[c]			.34		2.25**	1.30*
Meta-analytic weighted mean *g* (for all 4 studies)			**.87****		**1.80*****	**1.33*****

Note: A positive Hedge's *g* indicates improved compliance compared to the pretest or to the child release condition, based on DSTAT (Johnson, 1989). TO: timeout.

[a]TO criterion #1: mean number of TO episodes during the test session with 30 maternal commands.

[b]TO criterion #2: percentage of children requiring "excessive timeouts" (> 6) before remaining in TO for the required duration.

[c]Compared to the weighted mean for the child release condition from Bean and Roberts (1981) and Roberts and Powers (1990).

*$p < .05$.
**$p < .01$.
***$p < .001$.

This reflects a "Low" initial rating for non-randomized designs, positively adjusted for dose-gradient (i.e., linear association) and negatively adjusted for residual confounding and imprecision. Application of GRADE to back-up spanking resulted in a rating of "Moderate" certainty, more clearly for compliance with timeout than for compliance with parental commands (Table S-5). This grade reflects a High initial rating for randomized designs, negatively adjusted for imprecision and inconsistency (for one of the two compliance outcomes), and positively adjusted for large effects.

Discussion

Our primary overarching goal was to reconcile the results of Heilmann et al. (2021) narrative review with the two published meta-analyses of

controlled prospective longitudinal studies of child outcomes of physical punishment (Ferguson, 2013; Larzelere, Gunnoe, et al., 2018). By limiting our re-analysis to the studies available in 2021, we kept the focus on the relationship between methodological approaches (meta-analytic vs. narrative review; beta vs. slope) and the contrasting conclusions of the three previous literature reviews. (Readers interested in studies published after early 2020 are invited to view Supplementary Tables S-6 & S-7. Effect sizes from datasets/waves with initial analyses published after January 2020 are consistent with the mean effect sizes reported in Table 2.)

Prior results replicated

Our results make it quite easy to reconcile the results of the three published reviews. Consistent with the narrative review, almost half of the two-occasion intervals included in Table 2 yielded statistically significant *harmful*-looking effects of customary spanking based on the commonly used beta method (i.e., ANCOVA analyses). This was expected (and hypothesized) because the beta method is known to be biased against corrective actions (Berry & Willoughby, 2017; Hamaker et al., 2015; Hoffman, 2015; Larzelere, Lin, et al., 2018). Only two intervals yielded a significant beneficial-looking effect.

Consistent with the two meta-analyses, the mean effect sizes for the beta method were trivial, explaining 0.16% to 0.64% of the variance in four subsequent outcomes after controlling for initial scores on those outcomes. Trivial effect sizes like these can be accounted for by residual confounding, or by systemic noise, sometimes called a "crud factor" (Ferguson & Heene, 2021). It is likely that oppositional defiance creates a stronger confound of spanking with externalizing than with the other child outcomes, which may help explain the slightly larger meta-analytic beta obtained for externalizing (.08) relative to the other three outcomes (.04).

In contrast to results from the beta method, results from the slope method indicated more statistically significant *beneficial*-looking effects than harmful-looking ones. This too was *expected* (and hypothesized) because the slope method can overestimate the benefits of corrective actions due to regression toward the mean. Again, the effects were trivial in size. The explained variance for externalizing problems in the slope analyses (.09%) is also likely attributable to residual confounding or crud.

This pattern of tiny apparent effects, mostly in the direction of the biases associated with the methods used to produce them, replicates the reported results of all three published reviews of controlled longitudinal studies. It does not, however, lend equal support to their authors'

conclusions, which we will evaluate after the discussion of our specific goals.

Effects of customary spanking

Specific goal #1 was to estimate true average effect sizes for the causal impact of customary spanking on multiple child outcome, by contrasting effect sizes based on the beta method with effect sizes from analyses that are less biased than the beta method. Our pattern of results suggests that the true average causal effect of customary spanking on common child outcomes is near zero. We assert this for three reasons.

First, the most rigorous analyses we know how to conduct on available intervals of controlled longitudinal studies failed to yield any meta-analytic effect sizes that met/exceeded our SESOI. (Recall that prior to computing our estimates, we specified a SESOI – i.e., a smallest effect size of interest – of β = .10. We did this to prevent false positives (enabled by large sample sizes) based on effects that are so small that they could be caused by anything (Ferguson & Heene, 2021.) *All* eight of our average effect sizes in Table 2 were below the SESOI, meaning that we could not explain even 1% of the variance in either direction for any outcome using two different methods (beta and slope).

Second, the trivial estimates we obtained using the two different methods were in different directions. Per Angrist and Pischke (2009), estimates obtained from the two methods could be viewed as *bracketing* the true causal effect, given some statistical assumptions. (More simply, a good guess of the true causal effect would be halfway between the estimates.) The halfway values between the coefficients we obtained for externalizing problems range from .00 at 18 months to .07 at ages 10 and 11, explaining 0% to 0.5% of the variance in subsequent changes in externalizing problems. The halfway values of the mean beta and slope coefficients for the other three outcomes were .015, explaining barely more than two hundredths of one percent of the unexplained variance (0.0225%).

Third, our results were tested with two sensitivity tests on externalizing problems and one on cognitive achievement. Although the results of the two tests afforded little new information (to bolster or undercut our assertion), the fact that externalizing explicitly "failed" the backwards test suggests that the association between spanking and externalizing is *very* likely attributable to residual confounding or crud.

These current results – suggesting that the *true average* causal effects of customary spanking on four indicators of child adjustment are near zero – replicate the results of the prior (dual-method) MA that focused only on externalizing problems (Larzelere, Gunnoe, et al., 2018).

Tests of moderators

An *average* near-zero effect does not mean that the actual effect of each spanking is near-zero *for all, or even most, children*. It could mean that about half of spankings have a positive influence and the other half a negative influence. We therefore articulated specific goal #2: to identify situational and methodological factors associated with better- or worse-looking outcomes than the average effects of customary physical punishment.

Of 13 factors tested, the only factor that emerged as a moderator of externalizing problems was the one we had hypothesized: children's age. Again, results varied by the type of analyses conducted. Tests of moderation based on the *beta* method suggested a linear increase in the risk of spanking as children's age increased from 1.5 to 11 years. Tests of moderation based on the *slope* method suggested that spanking was associated with slight reductions in externalizing problems from 1.5 to 7 years of age, followed by slight increases from 8 to 11 years. This pattern replicated prior meta-analytic findings (Ferguson, 2013; Larzelere, Gunnoe, et al., 2018) that spanking looks more beneficial/less harmful during preschool than middle childhood, although curvilinear relationships should also be explored for the frequency and severity of spanking per Pritsker (2021) in combination with child temperament and age.

Unfortunately, our investigation of moderators was less comprehensive than we would have liked. We had an insufficient number of intervals to test for moderators of outcomes other than externalizing, and we had insufficient cases per response category to test some of the categorical moderators of externalizing behaviors that interested us the most (e.g., openhanded spanking vs. not; no spanking vs. an intermediate level of spanking).

Effects of back-up spanking

Specific goal #3 was to estimate unbiased average effect sizes for back-up spanking (i.e., two-swat openhanded spanking when a child refuses to cooperate with timeout) in clinical trials. These estimates indicated that children learned to cooperate with timeout and with parental commands more quickly with the spank back-up than in the control condition, which consisted of letting children decide when to end timeout.

We are aware that some researchers opposed to any spanking have attempted to discredit the relevance of these randomized trials (see Gershoff et al., 2018; Gershoff & Grogan-Kaylor, 2016). One assertion has been that the difference between post-test compliance rates for the first study listed in Table 3 were attributable to differing pretest rates. For this reason, we included the pretest rates of compliance in Table 3 – so that readers

could see that the pretest rates for the two groups were almost identical in the first study (23.3% vs. 23.4%).

A second assertion has been that parents do not need to spank because a brief room isolation was just as effective as the traditional spank back-up in enforcing cooperation with timeout in the second, third and fourth studies listed in Table 3. Room isolation is a condition outside the scope of the current MA, but a literature review of timeout variations concluded that *both* room isolation and spank contingencies were the most effective escape-prevention methods, based on these three randomized trials by Roberts (Everett et al., 2010, p. 247). We readily acknowledge the seemingly *equivalen*t average utility of both back-up methods, stressing that a lack of significance between the two most effective alternative treatments constitutes empirical evidence *for* a treatment, not evidence against it – according to the APA's criteria for empirical support for psychotherapies (Kaminski & Claussen, 2017). In the medical field, when two medications are deemed equally effective on average, both are kept available as treatment options. Having options is particularly helpful when there is evidence that some clients respond better to one type of treatment than another (a situation reported by the authors of these randomized studies: Roberts & Powers, 1990). Options are also necessary when use of one treatment is not appropriate for a particular situation. (From the beginning, Day & Roberts, 1983, questioned both the feasibility and ethics/safety of parental dependence on the room isolation, noting that many homes do not lend themselves to the construction or maintenance of an empty room of appropriate size with a half-height door to permit monitoring.)

Prior conclusions evaluated

As already stated, the results of the present MA replicate the findings of the narrative review by Heilmann et al. (2021) and the two relevant MAs of controlled longitudinal studies, even though the authors of these reviews came to very different conclusions. According to Heilmann and her team, their review constituted evidence "that physical punishment is harmful to children's development and wellbeing" and should be prohibited "in all forms and all settings" (p. 361). The results of the present MA are in keeping with the more nuanced conclusions proffered in the two relevant meta-analyses: the average effects of customary spanking are near zero. Thus, limited spanking (e.g., the "two-swat" protocol formerly taught in parent behavioral training programs) should continue to be an option for parents who need something other than room isolation to back up milder disciplinary consequences.

The pattern of evidence for customary spanking is similar to the pattern for most corrective actions, whether implemented by parents (e.g., timeout) or professionals (e.g., psychotherapy: Larzelere, Lin, et al., 2018; Larzelere,

Reitman, et al., 2023). The bias in the standard beta method may help explain why beta-based controlled longitudinal analyses of *alternative* disciplinary tactics have yet to identify a replacement for spanking that predicts significant improvements in child outcomes (Larzelere Cox, et al., 2010; Larzelere, Ferrer, et al., 2010; Larzelere, Knowles, et al., 2023).

State of parental discipline research

The state of child discipline research is not good. The vast majority of studies have been correlational or employed the beta-method. Longitudinal correlations make all corrective actions look harmful unless the prognosis of the problem is corrected *so perfectly* that the recipients of the action become indistinguishable from those who never had the problem at all (Larzelere et al., 2004). The beta method reduces this bias but fails to eliminate it, leaving residual confounding (Rothman et al., 2008). The slope method is not unbiased either, but it does remove the bias present in the beta method by adjusting fully for unchanging between-person differences that are already present in the initial outcome scores. *Only* randomized studies and equivalent designs can provide an unbiased estimate of the causal effect of a disciplinary technique, but there are only a handful of randomized studies of spanking, each based on a very small sample.

In addition to problems of research design, most studies of parental discipline lack information about the presenting problem and how the corrective disciplinary action was implemented. Evaluations of medical treatments routinely specify the presenting problem, the severity of the problem, the treatment dosage, and other details of treatment implementation. Researchers then compare alternative treatments *for a specified severity of a specified problem* (e.g., Glassburn et al., 1992), preferably with randomized comparisons. In contrast, most child discipline researchers ignore the presenting problem and ask imprecise questions about how the disciplinary response was implemented. They then proffer conclusions based on linear statistics (which are particularly influenced by the contrast between the minimum and maximum scores). In half of the datasets included in Table 2, the lowest possible score indicated an avoidance of spanking for one to four weeks (i.e., the absence of very recent spanking), and the highest possible score indicated at least three spankings per week (i.e., overly frequent recent spanking). This specifically contrasts overly frequent spanking with zero recent usage. Generalizing from these contrasts to claims that zero lifetime spanking leads to better outcomes than infrequent spanking is an inappropriate stretch (for evidence otherwise, see Gunnoe, 2013; Ellison et al., 2011).

Further inhibiting our knowledge of what constitutes effective discipline is the fact that many researchers are philosophically opposed to spanking, and some researchers now discourage *all* disciplinary consequences,

promoting only "positive parenting." (For a debate and empirical evaluations of positive parenting, see Holden et al., 2017, and Larzelere et al., 2017, 2020). This blanket opposition to all disciplinary consequences stifles follow-up evaluations that would be standard in the medical field. Medical researchers would likely have already conducted randomized trials to compare the effectiveness of back-up spanking, room isolation, positive parenting only, and other disciplinary responses for situations in which spanking was considered appropriate and effective in the past. The fact that these obvious research directions have never materialized may help explain why clinical treatments for child conduct problems are only half as effective now as when back-up spanking was prescribed (Weisz et al., 2019).

Spanking: future research directions

Longitudinal innovations recommended to overcome recently recognized biases in the beta method have documented non-trivial beneficial-looking effects of customary spanking (Berry & Willoughby, 2017, online Supporting Information, pp. 4-6) and of less frequent openhanded spanking (Pritsker, 2021). These innovations are a good start, but much more rigorous research is needed. Some specific questions needing more investigation are:

1. Is room isolation as effective as back-up spanking in homes (and clinics) that lack the ideal empty room employed in the original studies?
2. How quickly can back-up spanking (and alternatives) be *phased out* as the child learns to cooperate with nonphysical forms of discipline (e.g., Ellison et al., 2011; Gunnoe, 2013; Larzelere et al., 2013)?
3. What factors contribute to more vs. less appropriate uses of customary parental spanking (e.g., Lansford et al., 2012; Pritsker, 2021)? Some possible moderators include parent-child relatedness (biological vs. not), parent gender, the parent-child relationship (e.g., parenting style, attachment security), child characteristics (e.g., birth order, gender, temperament), community norms, mode of implementation (e.g., openhanded spanking vs. use of an instrument), the type of misbehavior eliciting the spanking, and relationship-restoring debriefing afterwards. (Although our results for fathers and various implementations, included in Table A1 in the Appendix, *generally* replicated the results in Table 2, they are based on very few studies. We need more data on fathers and various implementations.)
4. How do various types of spanking compare with each other and with alternative disciplinary responses? In the only meta-analysis to compare outcomes of physical discipline with alternative discipline tactics, Larzelere and Kuhn (2005) found that "conditional" spanking

> (conceptually similar to back-up spanking) was associated with significantly less aggression or noncompliance than 10 of 13 alternative tactics that it had been directly compared with. However, "overly severe" physical punishment was associated with more harmful-looking outcomes than alternative tactics. Thus, the weakness of the scientific evidence against customary spanking documented in this MA should *in no way* be used to justify harsh physical punishment which is opposed by virtually all contemporary discipline researchers.

Whenever possible, specific questions about discipline should be investigated in randomized trials, perhaps combined with regression discontinuity designs (Van Breukelen, 2013). This combination design could randomize undecided parents while assigning other parents to their preferred disciplinary technique. With instruction and oversight, this combination design could result in all conditions being more effective than the unsystematic "customary" versions of spanking or alternatives that parents would otherwise employ.

Research on child discipline could produce much more helpful recommendations by incorporating the rigorous designs and follow-up investigations that are routine in medical research. But well-designed studies of spanking can only happen if discipline researchers refrain from using trivial effect sizes from correlational and beta-based studies to depict customary spanking as so clearly detrimental that it is functionally "off the list" of corrective actions meriting rigorous empirical study.

Implications for parents

Learning to cooperate with adults is one of the primary socio-emotional tasks of early childhood. Parents who fail to foster this cooperative/compliant disposition in the early years put their children at risk for serious dysfunction during the child's adolescence and adulthood (Patterson & Fisher, 2002). Some children develop age-appropriate cooperation without being physically disciplined. Other children seem to benefit from the knowledge that defiance of parental directives (including the directive to abide milder forms of punishment) will elicit spanking. For some, the experience of age-delimited spanking in the context of a warm supportive parent-child relationship predicts positive outcomes.

Unfortunately, some well-intended researchers have emphasized methodologically weak/statistically-biased studies of physical discipline to convince parents that they should never spank. The current meta-analysis suggests that the harmful-effects of *customary* spanking have been exaggerated and the potential benefits of spanking for the most defiant children too quickly dismissed.

That said, there is still so much to be learned about the effects of spanking and alternatives in any particular situation. Because of this, we urge caution in the use of spanking, recommending that parents attempt to approximate, as much as possible, the conditions in which spanking was shown to be beneficial in clinical trials (i.e., two swats to the buttocks as a back-up when children aged 2–6 leave time-out prematurely). As demonstrated in these trials, parents who employed the judicious use of spanking as a back up to nonphysical punishments were often able to rapidly phase out the use of spanking. Rapid phase-out of spanking is a goal shared by all spanking researchers.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Professorship for Parenting Research, Oklahoma State University.

ORCID

Robert E. Larzelere http://orcid.org/0000-0003-2822-3735

References

Akcinar, B., & Baydar, N. (2016). Development of externalizing behaviors in the context of family and non-family relationships. *Journal of Child and Family Studies*, *25*(6), 1848–1859. https://doi.org/10.1007/s10826-016-0375-z

Altschul, I., Lee, S. J., & Gershoff, E. T. (2016). Hugs, not hits: Warmth and spanking as predictors of child social competence. *Journal of Marriage and the Family*, *78*(3), 695–714. https://doi.org/10.1111/jomf.12306

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's approach*. Princeton University Press. https://doi.org/10.1515/9781400829828

Ansari, A., & Gershoff, E. (2016). Parent involvement in Head Start and children's development: Indirect effects through parenting. *Journal of Marriage and Family*, *78*(2), 562–579. https://doi.org/10.1111/jomf.12266

Bakoula, C., Kolaitis, G., Veltsista, A., Gika, A., & Chrousos, G. P. (2009). Parental stress affects the emotions and behaviour of children up to adolescence: A Greek prospective, longitudinal study. *Stress*, *12*(6), 486–498. https://doi.org/10.3109/10253890802645041

Barajas-Gonzalez, R. G., Calzada, E., Huang, K.-Y., Covas, M., Castillo, C. M., & Brotman, L. M. (2018). Parent spanking and verbal punishment, and young child internalizing and externalizing behaviors in Latino immigrant families: Test of moderation by context and culture. *Parenting*, *18*(4), 219–242. https://doi.org/10.1080/15295192.2018.1524242

Barnes, J. C., Boutwell, B. B., Beaver, K. M., & Gibson, C. L. (2013). Analyzing the origins of childhood externalizing behavioral problems. *Developmental Psychology*, *49*(12), 2272–2284. https://doi.org/10.1037/a0032061

Baumrind, D., Larzelere, R. E., & Owens, E. B. (2010). Effects of preschool parents' power assertive patterns and practices on adolescent development. *Parenting*, *10*(3), 157–201. https://doi.org/10.1080/15295190903290790

Bean, A. W., & Roberts, M. W. (1981). The effect of time-out release contingencies on changes in child noncompliance. *Journal of Abnormal Child Psychology*, *9*(1), 95–105. https://doi.org/10.1007/BF00917860

Beauchaine, T. P., Webster-Stratton, C., & Reid, M. J. (2005). Mediators, moderators, and predictors of 1-year outcomes among children treated for early-onset conduct problems: A latent growth curve analysis. *Journal of Consulting and Clinical Psychology*, *73*(3), 371–388. https://doi.org/10.1037/0022-006X.73.3.371

Berger, L. M., Bruch, S. K., Johnson, E. I., James, S., & Rubin, D. (2009). Estimating the 'impact' of out-of-home placement on child well-being: Approaching the problem of selection bias. *Child Development*, *80*(6), 1856–1876. https://doi.org/10.1111/j.1467-8624.2009.01372.x

Berlin, L. J., Ispa, J. M., Fine, M. A., Malone, P. S., Brooks-Gunn, J., Brady-Smith, C., Ayoub, C., & Bai, Y. (2009). Correlates and consequences of spanking and verbal punishment for low-income White, African American, and Mexican American toddlers. *Child Development*, *80*(5), 1403–1420. https://doi.org/10.1111/j.1467-8624.2009.01341.x

Berry, D., & Willoughby, M. T. (2017). On the practical interpretability of cross-lagged panel models: Rethinking a developmental workhorse. *Child Development*, *88*(4), 1186–1206. https://doi.org/10.1111/cdev.12660

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley. https://doi.org/10.1002/9780470743386

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2011). *Comprehensive meta-analysis (Version 2.2)* [*Computer software*]. www.Meta-Analysis.com

Bornstein, M. H., Rothenberg, W. A., Bizzego, A., Bradley, R. H., Deater-Deckard, K., Esposito, G., Lansford, J. E., Putnick, D. L., & Zietz, S. (2023). Introduction and general methods: Parenting, national development, and early childhood development in 51 low- and middle-income countries. In M. H. Bornstein, W. A. Rothenberg, A. Bizzego, R. H. Bradley, K. Deater-Deckard, G. Esposito, J. E. Lansford, D. L. Putnick, & S. Zietz (Eds.), *Parenting and child development in low- and middle-income countries.* (pp. 1–51). Routledge. https://doi.org/10.4324/9781003044925-1

Bugental, D. B., Martorell, G. A., & Barraza, V. (2003). The hormonal costs of subtle forms of infant maltreatment. *Hormones and Behavior*, *43*(1), 237–244. https://doi.org/10.1016/S0018-506X(02)00008-9

Callender, K. A., Olson, S. L., Choe, D. E., & Sameroff, A. J. (2012). The effects of parental depressive symptoms, appraisals, and physical punishment on later child externalizing behavior. *Journal of Abnormal Child Psychology*, *40*(3), 471–483. https://doi.org/10.1007/s10802-011-9572-9

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* (2nd ed.). Lawrence Erlbaum.

Coley, R. L., Kull, M. A., & Carrano, J. (2014). Parental endorsement of spanking and children's internalizing and externalizing problems in African American and Hispanic families. *Journal of Family Psychology: Journal of the Division of Family Psychology of the American Psychological Association (Division 43)*, *28*(1), 22–31. https://doi.org/10.1037/a0035272

Cuartas, J., McCoy, D. C., Grogan-Kaylor, A., & Gershoff, E. (2020). Physical punishment as a predictor of early cognitive development: Evidence from econometric approaches. *Developmental Psychology*, *56*(11), 2013–2026. https://doi.org/10.1037/dev0001114

Day, D. E., & Roberts, M. W. (1983). An analysis of the physical punishment component of a parent training program. *Journal of Abnormal Child Psychology*, *11*(1), 141–152. https://doi.org/10.1007/BF00912184

Derella, O. J., Burke, J. D., Stepp, S. D., & Hipwell, A. E. (2020). Reciprocity in undesirable parent–child behavior? Verbal aggression, corporal punishment, and girls' oppositional defiant symptoms. *Journal of Clinical Child and Adolescent Psychology: The Official Journal for the Society of Clinical Child and Adolescent Psychology, American Psychological Association, Division 53*, *49*(3), 420–433. https://doi.org/10.1080/15374416.2019.1603109

Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental Psychology*, *50*(11), 2417–2425. https://doi.org/10.1037/a0037996

Ellison, C. G., Musick, M. A., & Holden, G. W. (2011). Does conservative Protestantism moderate the association between corporal punishment and child outcomes? *Journal of Marriage and Family*, *73*(5), 946–961. https://doi.org/10.1111/j.1741-3737.2011.00854.x

Everett, G. E., Hupp, S. D. A., & Olmi, D. J. (2010). Time-out with parents: A descriptive analysis of 30 years of research. *Education & Treatment of Children*, *33*(2), 235–259. https://doi.org/10.1353/etc.0.0091

Ferguson, C. J. (2013). Spanking, corporal punishment and negative long-term outcomes: A meta-analytic review of longitudinal studies. *Clinical Psychology Review*, *33*(1), 196–208. https://doi.org/10.1016/j.cpr.2012.11.002

Ferguson, C. J., & Heene, M. (2021). Providing a lower-bound estimate for psychology's "crud factor": The case of aggression. *Professional Psychology: Research and Practice*, *52*(6), 620–626. https://doi.org/10.1037/pro0000386

Font, S., & Cage, J. (2018). Dimensions of physical punishment and their association with children's cognitive performance and school adjustment. *Child Abuse & Neglect*, *75*, 29–40. https://doi.org/10.1016/j.chiabu.2017.06.008

Forehand, R. L., & McMahon, R. J. (1981). *Helping the noncompliant child*. Guilford Press.

Foshee, V. A., Ennett, S. T., Bauman, K. E., Benefield, T., & Suchindran, C. (2005). The association between family violence and adolescent dating violence onset. *The Journal of Early Adolescence*, *25*(3), 317–344. https://doi.org/10.1177/0272431605277307

Friedman, S. B., & Schonberg, S. K. (1996). Consensus statements [from the invitational conference, the short- and long-term consequences of corporal punishment]. *Pediatrics*, *98*(4), 852–853. https://doi.org/10.1542/peds.98.4.vi

Fu, R., Gartlehner, G., Grant, M., Shamliyan, T., Sedrakyan, A., Wilt, T. J., Griffith, L., Oremus, M., Raina, P., Ismaila, A., Santaguida, P., Lau, J., & Trikalinos, T. A. (2011). Conducting quantitative synthesis when comparing medical interventions: AHRQ and the effective health care program. *Journal of Clinical Epidemiology*, *64*(11), 1187–1197. https://doi.org/10.1016/j.clinepi.2010.08.010

Galton, F. (1886). Regression toward mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246–263. https://doi.org/10.2307/2841583

Gershoff, E. T. (2002). Corporal punishment by parents and associated child behaviors and experiences: A meta-analytic and theoretical review. *Psychological Bulletin*, *128*(4), 539–579. https://doi.org/10.1037/0033-2909.128.4.539

Gershoff, E. T., & Grogan-Kaylor, A. (2016). Spanking and child outcomes: Old controversies and new meta-analyses. *Journal of Family Psychology: Journal of the Division of Family Psychology of the American Psychological Association (Division 43)*, *30*(4), 453–469. https://doi.org/10.1037/fam0000191

Gershoff, E. T., Ansari, A., Purtell, K. M., & Sexton, H. R. (2016). Changes in parents' spanking and reading as mechanisms for Head Start impacts on children. *Journal of Family Psychology: Journal of the Division of Family Psychology of the American Psychological Association (Division 43)*, *30*(4), 480–491. https://doi.org/10.1037/fam0000172

Gershoff, E. T., Goodman, G. S., Miller-Perrin, C. L., Holden, G. W., Jackson, Y., & Kazdin, A. E. (2018). The strength of the causal evidence against physical punishment of children and its implications for parents, psychologists, and policymakers. *The American Psychologist*, *73*(5), 626–638. https://doi.org/10.1037/amp0000327

Gershoff, E. T., Lansford, J. E., Sexton, H. R., Davis-Kean, P., & Sameroff, A. J. (2012). Longitudinal links between spanking and children's externalizing behaviors in a national sample of White, Black, Hispanic, and Asian American families. *Child Development*, *83*(3), 838–843. https://doi.org/10.1111/j.1467-8624.2011.01732.x

Gershoff, E. T., Sattler, K. M. P., & Ansari, A. (2018). Strengthening causal estimates for links between spanking and children's externalizing behavior problems. *Psychological Science*, *29*(1), 110–120. https://doi.org/10.1177/0956797617729816

Gibson, C., & Fagan, A. (2018). An individual growth model analysis of childhood spanking on change in externalizing behaviors during adolescence: A comparison of Whites and African Americans over a 12-year-period. *American Behavioral Scientist*, *62*(11), 1463–1482. https://doi.org/10.1177/0002764218793689

Glassburn, J. R., Brady, L. W., Grigsby>., & B. W., Brady. (1992). Endometrium. In C. A. Perez & L. W(Eds.), *Principles and practice of radiation oncology.* (2nd ed., pp. 1203–1220). J. B. Lippincott.

Grogan-Kaylor, A. (2004). The effect of corporal punishment on antisocial behavior in children. *Social Work Research*, *28*(3), 153–162. https://doi.org/10.1093/swr/28.3.153

Grogan-Kaylor, A. (2005a). Corporal punishment and the growth trajectory of children's antisocial behavior. *Child Maltreatment*, *10*(3), 283–292. https://doi.org/10.1177/1077559505277803

Grogan-Kaylor, A. (2005b). Relationship of corporal punishment and antisocial behavior by neighborhood. *Archives of Pediatrics & Adolescent Medicine*, *159*(10), 938–942. https://doi.org/10.1001/archpedi.159.10.938

Grogan-Kaylor, A., Castillo, B., Ma, J., Ward, K. P., Lee, S. J., Pace, G. T., & Park, J. (2020). A Bayesian analysis of associations between neighborhoods, spanking and child externalizing behavior. *Children and Youth Services Review*, *112*, 104930. https://doi.org/10.1016/j.childyouth.2020.104930

Gromoske, A. N., & Maguire-Jack, K. (2012). Transactional and cascading relations between early spanking and children's social-emotional development. *Journal of Marriage and Family*, *74*(5), 1054–1068. https://doi.org/10.1111/j.1741-3737.2012.01013.x

Gunnoe, M. L. (2013). Associations between parenting style, physical discipline, and adjustment in adolescents' reports. *Psychological Reports*, *112*(3), 933–975. https://doi.org/10.2466/15.10.49.PR0.112.3.933-975

Gunnoe, M. L., & Mariner, C. L. (1997). Toward a developmental-contextual model of the effects of parental spanking on children's aggression. *Archives of Pediatrics & Adolescent Medicine*, *151*(8), 768–775. https://doi.org/10.1001/archpedi.1997.02170450018003

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*(1), 102–116. https://doi.org/10.1037/a0038889

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. https://doi.org/10.1002/jsrm.5

Heilmann, A., Mehay, A., Watt, R. G., Kelly, Y., Durrant, J. E., van Turnhout, J., & Gershoff, E. T. (2021). Physical punishment and child outcomes: A narrative review of prospective studies. *Lancet* , *398*(10297), 355–364. https://doi.org/10.1016/s0140-6736(21)00582-1

Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change*. Routledge.

Holden, G. W., Grogan-Kaylor, A., Durrant, J. E., & Gershoff, E. T. (2017). Researchers deserve a better critique: Response to Larzelere, Gunnoe, Roberts, and Ferguson (2017). *Marriage & Family Review*, *53*(5), 465–490. https://doi.org/10.1080/01494929.2017.1308899

Johnson, B. T. (1989). *DSTAT: Software for the meta-analytic review of research literatures.* Erlbaum.

Kaminski, J. W., & Claussen, A. H. (2017). Evidence base update for psychosocial treatments for disruptive behaviors in children. *Journal of Clinical Child and Adolescent Psychology: The Official Journal for the Society of Clinical Child and Adolescent Psychology, American Psychological Association, Division 53*, *46*(4), 477–499. https://doi.org/10.1080/15374416.2017.1310044

Keyser, D., Ahn, H., & Unick, J. (2017). Predictors of behavioral problems in young children 3 to 9 years old: The role of maternal and child factors. *Children and Youth Services Review*, *82*, 149–155. https://doi.org/10.1016/j.childyouth.2017.09.018

Lahey, B. B., Van Hulle, C. A., Keenan, K., Rathouz, P. J., D'Onofrio, B. M., Rodgers, J. L., & Waldman, I. D. (2008). Temperament and parenting during the first year of life predict future child conduct problems. *Journal of Abnormal Child Psychology*, *36*(8), 1139–1158. https://doi.org/10.1007/s10802-008-9247-3

Laible, D., Davis, A., Karahuta, E., & Van Norden, C. (2019). Does corporal punishment erode the quality of the mother–child interaction in early childhood? *Social Development*, *29*(3), 674–688. https://doi.org/10.1111/sode.12427

Lansford, J. E., Criss, M. M., Dodge, K. A., Shaw, D. S., Pettit, G. S., & Bates, J. E. (2009). Trajectories of physical discipline: Early childhood antecedents and developmental outcomes. *Child Development*, *80*(5), 1385–1402. https://doi.org/10.1111/j.1467-8624.2009.01340.x

Lansford, J. E., Criss, M. M., Laird, R. D., Shaw, D. S., Pettit, G. S., Bates, J. E., & Dodge, K. A. (2011). Reciprocal relations between parents' physical discipline and children's externalizing behavior during middle childhood and adolescence. *Development and Psychopathology*, *23*(1), 225–238. https://doi.org/10.1017/s0954579410000751

Lansford, J. E., Wager, L. B., Bates, J. E., Dodge, K. A., & Pettit, G. S. (2012). Parental reasoning, denying privileges, yelling, and spanking: Ethnic differences and associations with child externalizing behavior. *Parenting, Science and Practice*, *12*(1), 42–56. https://doi.org/10.1080/15295192.2011.613727

Lansford, J. E., Wager, L. B., Bates, J. E., Pettit, G. S., & Dodge, K. A. (2012). Forms of spanking and children's externalizing behaviors. *Family Relations*, *61*(2), 224–236. https://doi.org/10.1111/j.1741-3729.2011.00700.x

Larzelere, R. E., & Kuhn, B. R. (2005). Comparing child outcomes of physical punishment and alternative disciplinary tactics: A meta-analysis. *Clinical Child and Family Psychology Review*, *8*(1), 1–37. https://doi.org/10.1007/s10567-005-2340-z

Larzelere, R. E., Cox, R. B., Jr., & Mandara, J. (2013). Responding to misbehavior in young children: How authoritative parents enhance reasoning with firm control. In R. E. Larzelere, A. S. Morris, & A. W. Harrist (Eds.), *Authoritative parenting: Synthesizing nurturance and discipline for optimal child development.* (pp. 89–111). American Psychological Association. https://doi.org/10.1037/13948-005

Larzelere, R. E., Cox, R. B., Jr., & Smith, G. L. (2010). Do nonphysical punishments reduce antisocial behavior more than spanking? A comparison using the strongest previous causal evidence against spanking. *BMC Pediatrics*, *10*(1), 10. https://doi.org/10.1186/1471-2431-10-10

Larzelere, R. E., Ferrer, E., Kuhn, B. R., & Danelia, K. (2010). Differences in causal estimates from longitudinal analyses of residualized versus simple gain scores: Contrasting

controls for selection and regression artifacts. *International Journal of Behavioral Development*, *34*(2), 180–189. https://doi.org/10.1177/0165025409351386

Larzelere, R. E., Gunnoe, M. L., & Ferguson, C. J. (2018). Improving causal inferences in meta-analyses of longitudinal studies: Spanking as an illustration. *Child Development*, *89*(6), 2038–2050. https://doi.org/10.1111/cdev.13097

Larzelere, R. E., Gunnoe, M. L., Ferguson, C. J., & Roberts, M. W. (2019). The insufficiency of the evidence used to categorically oppose spanking and its implications for families and psychological science: Comment on Gershoff et al. (2018). *The American Psychologist*, *74*(4), 497–499. https://doi.org/10.1037/amp0000461

Larzelere, R. E., Gunnoe, M. L., Roberts, M. W., & Ferguson, C. J. (2017). Children and parents deserve better parental discipline research: Critiquing the evidence for exclusively "positive" parenting. *Marriage & Family Review*, *53*(1), 24–35. https://doi.org/10.1080/01494929.2016.1145613

Larzelere, R. E., Gunnoe, M. L., Roberts, M. W., Lin, H., & Ferguson, C. J. (2020). Causal evidence for *exclusively positive parenting* and for timeout: Rejoinder to Holden, Grogan-Kaylor, Durrant, and Gershoff (2017). *Marriage & Family Review*, *56*(4), 287–319. https://doi.org/10.1080/01494929.2020.1712304

Larzelere, R. E., Knowles, S. J., Adkison-Johnson, C., Cox, R. B., Jr., Lin, H., & Mandara, J. (2023). Ethnic differences in the effects of five disciplinary tactics on subsequent externalizing behavior problems. *Marriage & Family Review*, *59*(8), 523–548. https://doi.org/10.1080/01494929.2023.2199732

Larzelere, R. E., Kuhn, B. R., & Johnson, B. (2004). The intervention selection bias: An underrecognized confound in intervention research. *Psychological Bulletin*, *130*(2), 289–303. https://doi.org/10.1037/0033-2909.130.2.289

Larzelere, R. E., Lin, H., Payton, M. E., & Washburn, I. J. (2018). Longitudinal biases against corrective actions. *Archives of Scientific Psychology*, *6*(1), 243–250. https://doi.org/10.1037/arc0000052

Larzelere, R. E., Reitman, D., Ortiz, C., & Cox, R. B.Jr (2023). Parental punishment: Don't throw out the baby with the bathwater. In C. L. Frisby, R. E. Redding, W. T. O'Donohue, & S. O. Lilienfeld (Eds.), *Ideological and political bias in psychology: Nature, scope, and solutions*. (pp. 561–584). Springer.

Lee, S. J., Altschul, I., & Gershoff, E. T. (2013). Does warmth moderate longitudinal associations between maternal spanking and child aggression in early childhood? *Developmental Psychology*, *49*(11), 2017–2028. https://doi.org/10.1037/a0031630

Lee, S. J., Altschul, I., & Gershoff, E. T. (2015). Wait until your father gets home? Mothers' and fathers' spanking and development of child aggression. *Children and Youth Services Review*, *52*, 158–166. https://doi.org/10.1016/j.childyouth.2014.11.006

Lee, S. J., Grogan-Kaylor, A., & Berger, L. M. (2014). Parental spanking of 1-year-old children and subsequent child protective services involvement. *Child Abuse & Neglect*, *38*(5), 875–883. https://doi.org/10.1016/j.chiabu.2014.01.018

Lee, S. J., Pace, G. T., Ward, K. P., Grogan-Kaylor, A., & Ma, J. (2020). Household economic hardship as a moderator of the associations between maternal spanking and child externalizing behavior problems. *Child Abuse & Neglect*, *107*, 104573. https://doi.org/10.1016/j.chiabu.2020.104573

Lee, S. J., Taylor, C. A., Altschul, I., & Rice, J. C. (2013). Parental spanking and subsequent risk for child aggression in father-involved families of young children. *Children and Youth Services Review*, *35*(9), 1476–1485. https://doi.org/10.1016/j.childyouth.2013.05.016

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Sage.

Ma, J., & Grogan-Kaylor, A. (2017). Longitudinal associations of neighborhood collective efficacy and maternal corporal punishment with behavior problems in early childhood. *Developmental Psychology*, *53*(6), 1027–1041. https://doi.org/10.1037/dev0000308

Ma, J., Grogan-Kaylor, A., & Klein, S. (2018a). Neighborhood collective efficacy, parental spanking, and subsequent risk of household child protective services involvement. *Child Abuse & Neglect*, *80*, 90–98. https://doi.org/10.1016/j.chiabu.2018.03.019

Ma, J., Grogan-Kaylor, A., & Lee, S. J. (2018b). Associations of neighborhood disorganization and maternal spanking with children's aggression: A fixed-effects regression analysis. *Child Abuse & Neglect*, *76*, 106–116. https://doi.org/10.1016/j.chiabu.2017.10.013

Ma, J., Grogan-Kaylor, A., & Lee, S. J. (2020). Does community violence exposure moderate the associations between maternal spanking and early child behavior problems? *Aggressive Behavior*, *46*(3), 210–219. https://doi.org/10.1002/ab.21882

MacKenzie, M. J., Nicklas, E., Brooks-Gunn, J., & Waldfogel, J. (2015). Spanking and children's externalizing behavior across the first decade of life: Evidence for transactional trocesses. *Journal of Youth and Adolescence*, *44*(3), 658–669. https://doi.org/10.1007/s10964-014-0114-y

MacKenzie, M. J., Nicklas, E., Waldfogel, J., & Brooks-Gunn, J. (2013). Spanking and child development across the first decade of life. *Pediatrics*, *132*(5), e1118–e1125. https://doi.org/10.1542/peds.2013-1227

MacKenzie, M. J., Nicklas, E., Waldfogel, J., & Brooks-Gunn, J. (2012). Corporal punishment and child behavioural and cognitive outcomes through 5 years of age: Evidence from a contemporary urban birth cohort study. *Infant and Child Development*, *21*(1), 3–33. https://doi.org/10.1002/icd.758

Maguire-Jack, K., Gromoske, A. N., & Berger, L. M. (2012). Spanking and child development during the first 5 years of life. *Child Development*, *83*(6), 1960–1977. https://doi.org/10.1111/j.1467-8624.2012.01820.x

McLoyd, V. C., & Smith, J. (2002). Physical discipline and behavior problems in African American, European American, and Hispanic children: Emotional support as a mediator. *Journal of Marriage and Family*, *64*(1), 40–53. https://doi.org/10.1111/j.1741-3737.2002.00040.x

Mendez, M., Durtschi, J., Neppl, T. K., & Stith, S. M. (2016). Corporal punishment and externalizing behaviors in toddlers: The moderating role of positive and harsh parenting. *Journal of Family Psychology: Journal of the Division of Family Psychology of the American Psychological Association (Division 43)*, *30*(8), 887–895. https://doi.org/10.1037/fam0000187

Methods Group of the Campbell Collaboration. (2019). *Methodological expectations of Campbell Collaboration intervention reviews: Conduct standards*. Campbell Collaboration. https://doi.org/10.4073/cpg.2016.3

Moeyaert, M., Ugille, M., Natasha Beretvas, S., Ferron, J., Bunuan, R., & Van den Noortgate, W. (2017). Methods for dealing with multiple outcomes in meta-analysis: A comparison between averaging effect sizes, robust variance estimation and multilevel meta-analysis. *International Journal of Social Research Methodology*, *20*(6), 559–572. https://doi.org/10.1080/13645579.2016.1252189

Morgan, P. L., Li, H., Cook, M., Farkas, G., Hillemeier, M. M., & Lin, Y. C. (2016). Which kindergarten children are at greatest risk for attention-deficit/hyperactivity and conduct disorder symptomatology as adolescents? *School Psychology Quarterly: The Official Journal of the Division of School Psychology, American Psychological Association*, *31*(1), 58–75. https://doi.org/10.1037/spq0000123

Mulvaney, M. K., & Mebert, C. J. (2007). Parental corporal punishment predicts behavior problems in early childhood. *Journal of Family Psychology: Journal of the Division of*

Family Psychology of the American Psychological Association (Division 43), *21*(3), 389–397. https://doi.org/10.1037/0893-3200.21.3.389

Neaverson, A., Murray, A. L., Ribeaud, D., & Eisner, M. (2020). A longitudinal examination of the role of self-control in the relation between corporal punishment exposure and adolescent aggression. *Journal of Youth and Adolescence*, *49*(6), 1245–1259. https://doi.org/10.1007/s10964-020-01215-z

O'Gara, J. L., Calzada, E. J., LaBrenz, C., & Barajas-Gonzalez, R. G. (2020). Examining the longitudinal effect of spanking on young Latinx child behavior problems. *Journal of Child and Family Studies*, *29*(11), 3080–3090. https://doi.org/10.1007/s10826-020-01818-x

Okuzono, S., Fujiwara, T., Kato, T., & Kawachi, I. (2017). Spanking and subsequent behavior problems in toddlers: A propensity score-matched, prospective study in Japan. *Child Abuse & Neglect*, *69*, 62–71. https://doi.org/10.1016/j.chiabu.2017.04.002

Olson, S. L., Lopez-Duran, N., Lunkenheimer, E. S., Chang, H., & Sameroff, A. J. (2011). Individual differences in the development of early peer aggression: Integrating contributions of self-regulation, theory of mind, and parenting. *Development and Psychopathology*, *23*(1), 253–266. https://doi.org/10.1017/S0954579410000775

Paolucci, E. O., & Violato, C. (2004). A meta-analysis of the published research on the affective, cognitive, and behavioral effects of corporal punishment. *The Journal of Psychology*, *138*(3), 197–221. https://doi.org/10.3200/JRLP.138.3.197-222

Patterson, G. R., & Fisher, P. A. (2002). Recent developments in our understanding of parenting: Bidirectional effects, causal models, and the search for parsimony. In M. H. Bornstein (Ed.), *Handbook of parenting: Vol. 5. Practical issues in parenting.* (2nd ed., pp. 59–88). Erlbaum.

Petts, R. J., & Kysar-Moon, A. E. (2012). Child discipline and conservative Protestantism: Why the relationship between corporal punishment and child behavior problems may vary by religious context. *Review of Religious Research*, *54*(4), 445–468. https://doi.org/10.1007/s13644-012-0080-3

Piché, G., Huỳnh, C., Clément, M., & Durrant, J. E. (2017). Predicting externalizing and prosocial behaviors in children from parental use of corporal punishment. *Infant and Child Development*, *26*(4), 1–18. https://doi.org/10.1002/icd.2006

Pizer, S. D. (2016). Falsification testing of instrumental variables methods for comparative effectiveness research. *Health Services Research*, *51*(2), 790–811. https://doi.org/10.1111/1475-6773.12355

Pritsker, J. (2021). Spanking and externalizing problems: Examining within-subject associations. *Child Development*, *92*(6), 2595–2602. https://doi.org/10.1111/cdev.13701

Pustejovsky, J. E., & Tipton, E. (2017). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, *36*(4), 672–683. https://doi.org/10.1080/07350015.2016.1247004

Reeves, B. C., Deeks, J. J., Higgins, J. P. T., Shea, B., Tugwell, P., & Wells, G. A. (2019). Including non-randomized studies on intervention effects. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.), *Cochrane handbook for systematic reviews of interventions.* (2nd ed., pp. 565–620). Wiley-Blackwell. https://doi.org/10.1002/9781119536604.ch24

Reitman, D., & McMahon, R. J. (2013). Constance "Connie" Hanf (1917-2002): The mentor and the model. *Cognitive and Behavioral Practice*, *20*(1), 106–116. https://doi.org/10.1016/j.cbpra.2012.02.005

Roberts, M. W. (1988). Enforcing chair timeouts with room timeouts. *Behavior Modification*, *12*(3), 353–370. https://doi.org/10.1177/01454455880123003

Roberts, M. W., & Powers, S. W. (1990). Adjusting chair timeout enforcement procedures for oppositional children. *Behavior Therapy*, *21*(3), 257–271. https://doi.org/10.1016/S0005-7894(05)80329-6

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern epidemiology.* (3rd ed.). Wolter Kluwer.

Schunemann, H. J., Higgins, J. P. T., Vist, G. E., Glasziou, P., Akl, E. A., Skoetz, N., & Guyatt, G. H, (2019). Completing "summary of findings" tables and grading the certainty of the evidence. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. Welch (Eds.), *Cochrane handbook for systematic reviews of interventions.* (2nd ed., pp. 375–402). Wiley-Blackwell. https://doi.org/10.1002/9781119536604.ch14

Scott, S., Lewsey, J., Thompson, L., & Wilson, P. (2014). Early parental physical punishment and emotional and behavioural outcomes in preschool children. *Child: Care, Health and Development*, *40*(3), 337–345. https://doi.org/10.1111/cch.12061

Sege, R. D., Siegel, B. S., Flaherty, E. G., Gavril, A. R., Idzerda, S. M., Laskey, A. ", Legano, L. A., Leventhal, J. M., Lukefahr, J. L., Yogman, M. W., Baum, R., Gambon, T. B., Lavin, A., Mattson, G., Montiel-Esparza, R., & Wissow, L. S, (2018). Effective discipline to raise healthy children. *Pediatrics*, *142*(6), e20183112. https://doi.org/10.1542/peds.2018-3112

Slack, K. S., Holl, J. L., McDaniel, M., Yoo, J., & Bolger, K. (2004). Understanding the risks of child neglect: An exploration of poverty and parenting characteristics. *Child Maltreatment*, *9*(4), 395–408. https://doi.org/10.1177/1077559504269193

Slade, E. P., & Wissow, L. S. (2004). Spanking in early childhood and later behavior problems: A prospective study of infants and young toddlers. *Pediatrics*, *113*(5), 1321–1330. https://doi.org/10.1542/peds.113.5.1321

Stacks, A. M., Oshio, T., Gerard, J., & Roe, J. (2009). The moderating effect of parental warmth on the association between spanking and child aggression: A longitudinal approach. *Infant and Child Development*, *18*(2), 178–194. https://doi.org/10.1002/icd.596

Straus, M. A., & Paschall, M. J. (2009). Corporal punishment by mothers and development of children's cognitive ability: A longitudinal study of two nationally representative age cohorts. *Journal of Aggression, Maltreatment & Trauma*, *18*(5), 459–483. https://doi.org/10.1080/10926770903035168

Straus, M. A., Sugarman, D. B., & Giles-Sims, J. (1997). Spanking by parents and subsequent antisocial behavior of children. *Archives of Pediatrics & Adolescent Medicine*, *151*(8), 761–767. https://doi.org/10.1001/archpedi.1997.02170450011002

Taylor, C. A., Manganello, J. A., Lee, S. J., & Rice, J. C. (2010). Mothers' spanking of 3-year-old children and subsequent risk of children's aggressive behavior. *Pediatrics*, *125*(5), e1057–e1065. https://doi.org/10.1542/peds.2009-2678

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, *20*(3), 375–393. https://doi.org/10.1037/met0000011

Turns, B. A., & Sibley, D. S. (2018). Does maternal spanking lead to bullying behaviors at school? A longitudinal study. *Journal of Child and Family Studies*, *27*(9), 2824–2832. https://doi.org/10.1007/s10826-018-1129-x

Van Breukelen, G. J. P. (2013). ANCOVA versus CHANGE from baseline in nonrandomized studies: The difference. *Multivariate Behavioral Research*, *48*(6), 895–922. https://doi.org/10.1080/00273171.2013.831743

Wang, M. T., & Kenny, S. (2014). Parental physical punishment and adolescent adjustment: Bidirectionality and the moderation effects of child ethnicity and parental warmth. *Journal of Abnormal Child Psychology*, *42*(5), 717–730. https://doi.org/10.1007/s10802-013-9827-8

Ward, K. P., Lee, S. J., Limb, G. E., & Grogan-Kaylor, A. C. (2021). Physical punishment and child externalizing behavior: Comparing American Indian, White, and African American children. *Journal of Interpersonal Violence*, *36*(17–18), NP9885–NP9907. https://doi.org/10.1177/0886260519861678

Ward, K. P., Lee, S. J., Pace, G. T., Grogan-Kaylor, A., & Ma, J. (2020). Attachment style and the association of spanking and child externalizing behavior. *Academic Pediatrics*, *20*(4), 501–507. https://doi.org/10.1016/j.acap.2019.06.017

Weisz, J. R., Kuppens, S., Ng, M. Y., Vaughn-Coaxum, R. A., Ugueto, A. M., Eckshtain, D., & Corteselli, K. A. (2019). Are psychotherapies for young people growing stronger? Tracking trends over time for youth anxiety, depression, attention-deficit hyperactivity disorder, and conduct problems. *Perspectives on Psychological Science: a Journal of the Association for Psychological Science*, *14*(2), 216–237. https://doi.org/10.1177/1745691618805436

Wells, G. A., Shea, B., Higgins, J. P. T., Sterne, J., Tugwell, P., & Reeves, B. C. (2013). Checklists of methodological issues for review authors to consider when including non-randomized studies in systematic reviews. *Research Synthesis Methods*, *4*(1), 63–77. https://doi.org/10.1002/jrsm.1077

Xing, X., Wang, M., Zhang, Q., He, X., & Zhang, W. (2011). Gender differences in the reciprocal relationships between parental physical aggression and children's externalizing problem behavior in China. *Journal of Family Psychology journal of the Division of Family Psychology of the American Psychological Association (Division 43)*, *25*(5), 699–708. https://doi.org/10.1037/a0025015

Yoo, J. A., & Huang, C.-C. (2013). Long-term relationships among domestic violence, maternal mental health and parenting, and preschool children's behavior problems. *Families in Society: The Journal of Contemporary Social Services*, *94*(4), 268–276. https://doi.org/10.1606/1044-3894.4321

Yu, J., Cheah, C. S. L., Hart, C. H., & Yang, C. (2018). Child inhibitory control and maternal acculturation moderate effects of maternal parenting on Chinese American children's adjustment. *Developmental Psychology*, *54*(6), 1111–1123. https://doi.org/10.1037/dev0000517

Appendix

Table A1. Effect sizes from other qualifying controlled longitudinal studies and why they did not contribute to the main featured meta-analysis.

	Dataset Already Included[a]	Correlation Matrix Available?	β	$r_{x(y2-y1)}$
Externalizing Behavior Problems				
Barajas-Gonzalez et al. (2018)	NYC Hispanics	No	.03	
Berry and Willoughby (2017)				
FFCW data	FFCW 3yrs	No	.08**	−.10**
Gershoff et al. (2018)	ECLS-K 1998	No	.07**	
Gibson and Fagan (2018)	None	No	.06^	
Grogan-Kaylor et al. (2020)	FFCW 3yrs	No	.07***	
Laible et al. (2019)	SECCYD 54 mos.	Yes	−.26***[d]	−.24***[d]
Lansford et al. (2009)	CDP (replaced with 2011 study's correlation matrices)			
Lansford et al. (2009)	Pitt (replaced with 2011 study)			
Lansford et al. (2012, *Parenting*)	CDP 5 yrs[b]	Yes	.12**	.02
Lansford et al. (2012, *FamRel* 6yrs)	CDP[bc]	Yes	.03[d]	.00[d]
Lansford et al. (2012, *FamRel* 7yrs)	CDP[bc]	Yes	−.01[d]	−.02[d]
Lee et al. (2013)				
3+ times/mo. (mother)	FFCW 3yrs	No	.14**	
3+ times/mo. (father)	Fathers	No	.13*	
1–2 times/mo. (mother)	FFCW 3yrs[c]	No	.03	
1–2 times/mo. (father)	Fathers[c]	No	.04	
Lee et al. (2015)				
Mother	FFCW 3yrs	No	.10**	
Fathers	Fathers	No	(n.s.)	
Lee et al. (2020)				
From age 3 to 5	FFCW 3yrs	No	.03**	
From age 5 to 9	FFCW 5 yrs	No	.03**	
MacKenzie et al. (2012)				
2+ times/wk (mother)	FFCW 3yrs	No	.06*	
2+ times/wk (father)	Fathers	No	.04	
1 time/wk (mother)	FFCW 3yrs[c]	No	.02	
1 time/wk (father)	Fathers[c]	No	.05	
MacKenzie et al. (2013)				
2+ times/wk (mother)	FFCW 5 yrs	No	.06**	
2+ times/wk (father)	Fathers	No	.01	
1 time/wk (mother)	FFCW 5 yrs[c]	No	.05*	
1 time/wk (father)	Fathers[c]	No	.02	
MacKenzie et al. (2015)				
From age 3 to 5	FFCW 3yrs	No	.05*	
From age 5 to 9	FFCW 5 yrs	No	.05*	
McLoyd and Smith (2002)	None	No		−.06*
Mendez et al. (2016)	Fathers	Yes	.21**[d]	.12[d]
Morgan et al. (2016)	ECLS-K	No	.09*	
Petts and Kysar-Moon (2012)	FFCW 3yrs	No	.09**	
Stacks et al. (2009)	Early Head Start	No	.03	
Straus et al. (1997)	NLSY: 1986 3–5yrs	No	.08*[de]	−.06[de]
Straus et al. (1997)	NLSY: 1986 6+ yrs	No	.07*[de]	−.07*[de]
Straus et al. (1997)	NLSY: 1988 3–5yrs	No	.16***[de]	.03[de]
Taylor et al. (2010)				
3+ times/mo.	FFCW 3yrs	No	.11***	
1–2 times/mo.	FFCW 3yrs[c]	No	.04	
Turns and Sibley (2018)	FFCW 3 yrs	Yes	.08	−.08^
Turns and Sibley (2018)	FFCW 5 yrs	Yes	.07	.03

(*Continued*)

Table A1. Continued.

	Dataset Already Included[a]	Correlation Matrix Available?	β	$r_{x(y2\text{-}y1)}$
Ward et al. (2020)	FFCW 3yrs	No	.07**	
Ward et al. (2021)	FFCW 3yrs	No	.08***	
Unweighted Means for Externalizing Behavior Problems			**.06*** **	**−.04*** **
Internalizing Behavior Problems				
Barajas-Gonzalez et al. (2018)	NYC Hispanic	No	.01	
Petts and Kysar-Moon (2012)	FFCW 3yrs	No	.11***	
Unweighted Mean for Internalizing Behavior Problems			**.06***	
Cognitive Achievement				
MacKenzie et al. (2012)				
2+ times/wk (mother)	FFCW 3yrs	No	−.06^	
2+ times/wk (father)	FFCW 3yrs	No	.04	
1 time/wk (mother)	FFCW 3yrs	No	.01	
1 time/wk (father)	FFCW 3yrs	No	.04	
MacKenzie et al. (2013)				
2+ times/wk (mother)	FFCW 5yrs	No	−.02	
2+ times/wk (father)	FFCW 5yrs	No	−.07^	
1 time/wk (mother)	FFCW 5yrs	No	.01	
1 time/wk (father)	FFCW 5yrs	No	−.03	
Unweighted Mean for Cognitive Achievement			**−.01**	

[a]FFCW: Fragile Families and Child Well-Being; ECLS-K: Early Childhood Longitudinal Study—Kindergarten Cohort 1998–1999; SECCYD: Study of Early Child Care and Youth Development; NLSY: National Longitudinal Survey of Youth.
[b]Openhanded spanking.
[c]Includes test of intermediate spanking frequency.
[d]Beta and slope coefficients based on correlation matrix.
[e]Effect sizes from Straus et al. (1997) were estimated using their reported correlates of spanking from each cohort, along with the stability coefficient for antisocial behavior from the attempted duplication by Larzelere et al. (2010) for 6- to 9-year-olds in 1988.
^$p < .10$.
*$p < .05$.
**$p < .01$.
***$p < .001$.